Review of U.S. EPA's ORD Staff Handbook for Developing IRIS Assessments: 2020 Version

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

Review of U.S. EPA's ORD Staff Handbook for Developing IRIS Assessments: 2020 Version

Committee to Review EPA'S IRIS Assessment Handbook

Board on Environmental Studies and Toxicology

Division on Earth and Life Studies

A Consensus Study Report of
The National Academies of
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by Contract No. 68HERC19D0011 between the National Academy of Sciences and the U. S. Environmental Protection Agency. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-XXXXX-X
International Standard Book Number-10: 0-309-XXXXX-X
Digital Object Identifier: https://doi.org/10.17226/26289

Additional copies of this publication are available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2021. *A Review of U.S. EPA's ORD Staff Handbook for Developing IRIS Assessments: 2020 Version*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26289.

The National Academies of

SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. John L. Anderson is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

COMMITTEE TO REVIEW EPA'S IRIS ASSESSMENT HANDBOOK

Members

Lisa A. Bero *(Chair)*, Colorado School of Public Health, Aurora
Hugh A. Barton, Consultant, Mystic, CT
Weihsueh A. Chiu, Texas A&M University, College Station
Gary L. Ginsberg, New York State Department of Health, Albany
Julie B. Herbstman, Columbia University, New York, NY
Jessica L. Myers, Texas Commission on Environmental Quality, Austin
Heather B. Patisaul, North Carolina State University, Raleigh
David B. Richardson, University of North Carolina, Chapel Hill
Paul Whaley, Lancaster University, Lancaster, United Kingdom
Kristine L. Willett, University of Mississippi, University
Corwin M. Zigler, University of Texas, Austin

Staff

Raymond Wassel, Scholar
Clifford S. Duke, Director, Board on Environmental Studies and Toxicology
Kathryn Guyton, Senior Program Officer
Leslie Beauchamp, Program Assistant
Melissa Franks, Senior Program Officer (until May 14, 2021)

Sponsor

U.S. Environmental Protection Agency

BOARD ON ENVIRONMENTAL STUDIES AND TOXICOLOGY

Members

William H. Farland (*Chair*), Colorado State University, Fort Collins
Dana Boyd Barr, Emory University, Atlanta, GA
Ann M. Bartuska, U.S. Department of Agriculture (retired), Washington, DC
E. William Colglazier, American Association for the Advancement of Science, Washington, DC
Francesca Dominici, Harvard University, Boston, MA
George Gray, The George Washington University, Washington, DC
R. Jeffrey Lewis, ExxonMobil Biomedical Sciences, Inc., Annandale, NJ
Germaine M. Buck Louis, George Mason University, Fairfax, VA
Linsey C. Marr, Virginia Polytechnic Institute and State University, Blacksburg
R. Craig Postlewaite, U.S. Department of Defense, Burke, VA
Reza J. Rasoulpour, Corteva Agriscience, Indianapolis, IN
Ivan Rusyn, Texas A&M University, College Station
Joshua Tewksbury, Future Earth, Boulder, CO
Sacoby M. Wilson, University of Maryland, College Park

Staff

Clifford S. Duke, Director
Raymond Wassel, Scholar
Kathryn Guyton, Senior Program Officer
Kaley Beins, Associate Program Officer
Laura Llanos, Finance Business Partner
Tamara Dawson, Program Coordinator
Leslie Beauchamp, Program Assistant
Thomasina Lyles, Program Assistant

Acknowledgments

This Consensus Study Report was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each published report as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their review of this report:

HARVEY CLEWELL, Ramboll US Consulting, Inc.
PENNY FENNER-CRISP, private consultant
SUE JINKS-ROBERTSON (NAS), Duke University
JULEEN LAM, California State University East Bay
JENNIFER McPARTLAND, Environmental Defense Fund
MOSHE OREN (NAS/NAM), Weizmann Institute of Science
KAREN ROBINSON, Johns Hopkins University
ANDREW ROONEY, National Institute of Environmental Health Sciences
KAN SHAO, Indiana University
KIM TRUONG, California Environmental Protection Agency
LAURA VANDENBERG, University of Massachusetts – Amherst

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations of this report nor did they see the final draft before its release. The review of this report was overseen by JOEL KAUFMAN, University of Washington and LYNN GOLDMAN (NAM), the George Washington University. They were responsible for making certain that an independent examination of this report was carried out in accordance with the standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the authoring committee and the National Academies.

Abbreviations and Acronyms

ADME	Absorption, Distribution, Metabolism, Excretion
BMR	Benchmark Response
CRO	Contract Research Organization
EPA	U.S. Environmental Protection Agency
GRADE	Grading of Recommendations, Assessment, Development and Evaluation
HAWC	Health Assessment Workspace Collaborative
IAP	IRIS Assessment Plan
IRIS	Integrated Risk Information System
KC	Key Characteristics
LOAEL	Lowest-observed-adverse-effect Level
MOA	Mode of Action
NAMs	New Approach Methods/Methodologies
OHAT	Office of Health Assessment and Translation
OPPT	U.S. Environmental Protection Agency Office of Pollution Prevention and Toxics
ORD	U.S. Environmental Protection Agency Office of Research and Development
PBPK	Physiologically Based Pharmacokinetic
PECO	Population, Exposure, Comparator, Outcome
PK	Pharmacokinetic
PODs	Points of Departure
RfC	Reference Concentration
RfD	Reference Dose
ROBINS-E	Risk of Bias in Non-randomized Studies of Exposures
ROBINS-I	Risk of Bias in Non-randomized Studies of Interventions
TACIT	Tool for Addressing Conflicts of Interest in Trials
TK	Toxicokinetics
TSCA	Toxic Substances Control Act

Contents

APPENDIXES

BOXES, FIGURES, AND TABLES

BOXES

FIGURES

TABLES

Summary

The U.S. Environmental Protection Agency's (EPA) Integrated Risk Information System (IRIS) program develops human health assessments that focus on hazard identification and dose-response analyses for chemicals in the environment. The assessments are highly important as they are routinely used to inform risk assessments and risk management decisions across the agency. In addition, they are used by federal, state, local, and tribal agencies, as well as community organizations and agencies in other countries, to inform decisions concerning health risk assessment and management.

Over the years, questions had been raised about the scientific basis of toxicity values reported in some IRIS assessments and the extensive amount of time taken to complete assessments. The 2014 report of the National Academies of Sciences, Engineering, and Medicine entitled *Review of EPA's Integrated Risk Information System (IRIS) Process* recommended that EPA develop a single handbook to provide detailed guidance for all those involved in developing IRIS assessments. The *ORD Staff Handbook for Developing IRIS Assessments* (the handbook) provides guidance to scientists who perform the IRIS assessments in order to foster consistency in the assessments and enhance transparency about the IRIS assessment process. It also provides an opportunity for stakeholder communities to become aware of the processes and policies guiding those who are drafting IRIS assessments. The handbook describes the tasks involved in carrying out the sequential stages in preparing an IRIS draft assessment (see Box S-1).

EPA requested that the National Academies conduct a review of the 2020 version of the handbook. In response, the Committee to Review EPA's IRIS Assessment Handbook has been convened to review the procedures and considerations for operationalizing the principles of systematic reviews and the methods described in the handbook for determining the scope of the IRIS assessments, evidence integration, extrapolation techniques, dose-response analyses, and characterization of uncertainties.

GENERAL COMMENTS ON THE HANDBOOK

The committee found that the handbook reflects the significant improvements that EPA has made in its IRIS assessment process. For instance, the handbook describes the inclusion of sophisticated, state-of-the-art methods that use systematic evidence maps to summarize literature characteristics for scoping and systematic review methods for hazard identification. Moreover, the IRIS program is clearly helping to advance the science of systematic review, as applied to hazard identification. EPA staff are actively involved in the ongoing development of methods, such as study evaluation and handling of mechanistic data. The committee recognizes that EPA faces challenges in implementing many of the methods for the IRIS assessment process and is impressed and encouraged by the progress that the IRIS program has made to date. The methods for developing IRIS assessments can serve as a model for other EPA programs that are implementing systematic review methods.

BOX S-1 Stages in the Development of IRIS Assessments Listed in the Handbook

Scoping to define the parameters of the assessment.

Problem formulation: Preliminary literature survey. Describe health effects of potential interest and key science issues including predefined mechanistic analyses. Develop an assessment plan containing draft population, exposure, comparator, outcome (PECO) criteria.

Systematic review protocol: Describe systematic review procedures for PECO criteria, literature identification, study evaluation, and data extraction/display.

Literature search and screening: Identify health effect studies and other informative studies relevant to evaluating potential health effects.

Literature inventory of human, animal, and mechanistic studies.

Refined evaluation plan: Prioritize and define endpoint groupings, refine PECO.

Study evaluation: Evaluate health effect studies for risk of bias and insensitivity; evaluate physiologically based pharmacokinetic modeling and simulation models and other information as needed.

Organize hazard review: Presentation decisions (i.e., tables and graphs). Organize and prioritize relevant mechanistic information.

Data extraction and display: Human and animal health effect studies.

Synthesis

- Human studies
- Animal studies
- Mechanistic information (data extraction, display, analysis, and synthesis)

Integration

- Summarize the strength of each evidence stream as part of the evidence integration narrative
- Overall evidence integration across evidence streams (hazard identification, including review of susceptibility)

Hazard consideration and study selection for deriving toxicity values.

Derive toxicity values: Cancer and/or noncancer.

However, the committee found that the handbook does not consistently convey the strengths and advances in methodology for the IRIS assessment process in an even and clear manner. Thus, the committee offers recommendations aimed at ensuring the handbook meets its objectives of providing transparency about the IRIS assessment process and providing operational instructions for those conducting the assessments. This summary provides the committee's highest priority recommendations that it believes are critical for improving the scientific rigor and clarity of the handbook. Other lower priority suggestions can be found in the conclusions of individual chapters of this report.

OVERVIEW OF ORGANIZATION AND CONTENT OF THE HANDBOOK

The overall organization and presentation of the handbook are in need of improvement. For example, there are multiple places in the handbook where the roles of mechanistic and toxicokinetic (TK)[1] data are described, but they are not entirely consistent. Similarly, multiple places in the handbook describe how the evidence base for susceptible populations should be handled. However, the definition of what constitutes evidence of susceptibility and the types of data that may inform such susceptibility are not defined in one place. In addition, the handbook chapter on selection of studies for toxicity value determination does not directly follow on from the earlier handbook chapters. As described, hazard evaluation and toxicity value determination appear to be disconnected processes.

> **Recommendation 2.1:** EPA should engage a professional editor with specific expertise in developing handbook-like materials to assist with the handbook revision. The editor should enhance the transparency and ease of use of the handbook by focusing the material on the concepts, definitions, and instructions needed to complete the main steps in the IRIS assessment process; eliminating unnecessary repetition among the chapters; and ensuring that terminology is used consistently across chapters.

The handbook uses some of its terminology in an inconsistent manner, a substantive issue. For instance, definitions for key terms such as "scoping" and "sensitivity" are currently scattered across different chapters of the handbook. In other cases, the handbook assigns unconventional definitions to terms used for designating various IRIS-specific processes and products in the area of evidence synthesis.

> **Recommendation 2.2:** EPA should add a glossary to the handbook for defining key terms. Single definitions should be provided for concepts, and the definitions should be applied consistently throughout the handbook.

> **Recommendation 2.3:** The handbook should use terminology in a manner that is consistent with existing, accepted definitions in related fields. When alternative definitions are used for the IRIS assessment process, the handbook should provide explicit justification.

The handbook does not clearly indicate where systematic review methods are used in the major steps of the IRIS assessment process and where they are not used.

> **Recommendation 2.4:** When systematic review methods are being used for parts of the IRIS assessment process, this should be stated and the relevant methodological literature should be referenced.

---

[1] In this report, toxicokinetics refers to the absorption, distribution, metabolism, and excretion processes also called ADME or pharmacokinetics. However, consistent with the handbook, models are described as pharmacokinetic or physiologically based pharmacokinetic models rather than toxicokinetic models.

The handbook does not adequately describe the overall process or flow for developing IRIS assessments and the iterative nature of some of the steps in the process.

> **Recommendation 2.5:** EPA should create new graphical and tabular depictions of the IRIS assessment process as it is currently practiced, and should not feel constrained to mirror the process depicted by the 2014 National Academies report *Review of EPA's Integrated Risk Information System (IRIS) Process* or any other report (including this one). A professional editor could be of assistance here.

The handbook includes very detailed information on some methods that may undergo rapid development (e.g., data extraction). On the other hand, the handbook lacks specific examples from relevant IRIS assessments and examples of software used by EPA, such as the Health Assessment Workspace Collaborative (HAWC), that are needed to fully understand the IRIS program's methods.

> **Recommendation 2.6:** The handbook should incorporate more examples from relevant IRIS assessments and examples of software used by EPA, such as HAWC. EPA could provide them as supplementary material or links to other content.

Funding bias refers to an association between study funding sources and financial ties of investigators with research outcomes that are favorable for the sponsors. Publication bias is the publication or non-publication of research results based on the nature and direction of the results. Publication bias and funding bias are mentioned sparingly in the handbook, although empirical evidence from a variety of fields shows that funding bias and publication bias can alter effect estimates when evidence is synthesized. Funding bias might also have an effect on the confidence of study ratings from evidence evaluation. Tools and methods to detect and assess these biases are available.

> **Recommendation 2.7:** The handbook should describe how to detect and assess the effect of funding bias on the confidence of study ratings from evidence evaluation or effect estimates from synthesis.

> **Recommendation 2.8:** The handbook should describe how to detect and assess the effect of publication bias on effect estimates from synthesis.

PLANNING THE IRIS ASSESSMENT

Chapters 1 – 7 and 10 of the handbook describe processes related to planning the development of IRIS assessments. The committee divided its critiques of the overall planning process for IRIS assessments into three general phases: problem formulation, protocol development, and organization of the hazard review.

Problem Formulation

The problem formulation process described in the handbook presents an important evolution in the understanding of how to use systematic evidence maps (also called "literature

inventories" in the handbook) to inform priority setting in toxicity assessments. The process of developing the systematic evidence maps is in accordance with established best practices. The problem formulation process makes suitable use of information specialists and is notable for its comprehensive coverage of both published and grey literature resources. The systematic evidence map is a key milestone in the IRIS process, as it incorporates information obtained from the scoping step and initial problem formulation step and forms the foundation of all subsequent analysis.

The committee notes that systematic evidence maps also have considerable potential value in and of themselves as a public good by providing a publicly accessible database that could be queried and used by any research organization to identify knowledge gaps and clusters in toxicity research.

> **Recommendation 3.1:** The handbook should make explicit which components of a literature inventory database are to be made publicly available and when.

Protocol Development

The handbook lacks clarity regarding the products of the planning process, the relationships among them, which are expected to be updated or should be registered, and how they feed into the IRIS Assessment. In a systematic review, the protocol is a complete account of planned methods, which should be registered prior to conduct of the review. The term "registration," in this context, is generally understood to mean the public release of the protocol in a time-stamped, read-only format. The handbook lacks clarity as to exactly what documented output of the assessment process constitutes a protocol, as the term "protocol" seems to refer to as many as three types of documents. The scope of each of these documents is also described ambiguously in the handbook.

> **Recommendation 3.5:** The handbook should clarify and simplify the assessment planning process as follows: restructure the handbook to directly reflect the order in which each step is undertaken, unambiguously identify each of the products of the planning process, clearly define what each product consists of, and state if and when each product is to be made publicly available.

> **Recommendation 3.6:** EPA should create a time-stamped, read-only, final version of each document that details the planned methods for an IRIS assessment prior to conducting the assessment.

The committee observes that the processes for inclusion and exclusion of studies from the systematic reviews conducted as part of the IRIS assessments diverges from current best practices. In conventional systematic reviews, the inclusion and exclusion of evidence, as well as the delineation of "units of analysis" for evidence synthesis, are strictly governed by prespecified population, exposure, comparator, outcome (PECO) statements. However, the handbook, as well as recently released IRIS assessment planning documents (e.g., for Vanadium and Inorganic Mercury Salts), only include broad PECO statements, along with broad health effect categories. This approach contrasts with conventional systematic reviews, in which even within a relatively broad health effect category (e.g., cardiovascular disease), a detailed PECO statement with

specific outcomes is prespecified to be fully transparent and to minimize potential for bias via selective inclusion of literature in a review.

Moreover, as currently described in the handbook, evidence may be excluded from further consideration at the "refined plan" step or at the "organize hazard review" step. The reasons for the triage of such evidence, and safeguards to ensure that the evidence is not being used selectively, are not explained sufficiently in the handbook.

> **Recommendation 3.7:** The IRIS assessment protocol should include refined PECO statements for each unit of analysis defined at the levels of endpoint or health outcome. The development of the refined PECO statements could benefit from considerations of available mechanistic data (e.g., grouping together causally linked endpoints, separating animal evidence by species or strain) and TK information (e.g., grouping or separating evidence by route of exposure), and information about population susceptibility.

Organization of the Hazard Review

While there should be explicit consideration given to the organization of the hazard review in the planning stages of the IRIS assessment, the handbook leaves this organizing process under-specified. This is problematic given that this stage requires the highest level of granularity in describing how evidence is to be selected, organized, and grouped prior to the data extraction stage of an assessment. Much more clarity is needed about which elements would more appropriately be subsumed under other (earlier) stages. Moreover, this step deviates from best practices in conventional systematic reviews, where all outcomes are prespecified and not subject to change after study evaluation. The handbook does not provide sufficient justification for revisiting the design of the systematic review after study evaluation. In particular, it does not clearly articulate what findings from the study evaluation would be sufficient to change the analysis plan.

> **Recommendation 3.8:** The steps of organizing the hazard review should be narrowed to focus on new information obtained after the study evaluation stage. Organizing the hazard review should be structured by clear criteria for triage and prioritization, and aimed at producing transparent documentation of how and why outcomes and measures are being organized for synthesis.

The Use of Mechanistic and Toxicokinetic Data and Key Characteristics

The roles of mechanistic and TK data in the planning process are described in multiple places, and the descriptions are not entirely consistent. In some cases when it is appropriate and possible to evaluate the strength of evidence of mechanistic and TK data or pharmacokinetic (PK) models using systematic review methods, these data may warrant their own PECO (or a set of PECO) statements. However, the handbook is not clear as to when mechanistic and TK data or PK models may require a separate PECO statement defining a discrete unit of analysis for systematic review, synthesis, and strength of evidence judgments.

Key characteristics (KCs) comprise the set of chemical and biological properties of agents that cause a particular toxic outcome. Although the use of KCs to search, screen, and organize mechanistic data is increasingly becoming accepted, the role of KCs for informing

hazard identification has been the subject of debate. They are appropriate for use in evaluating biological plausibility, or lack thereof. However, KCs as currently constructed tend to be sensitive, but not necessarily specific. More research is needed into whether and how they can be used to be more predictive of hazard.

> **Recommendation 3.9:** The handbook should describe how the IRIS assessment plan and IRIS assessment protocol can identify the potential roles of mechanistic and TK data, including if they are to be units of analysis for systematic review, synthesis, and strength of evidence judgments. At a minimum, all endpoints that may be used for toxicity values, including so-called "precursor" endpoints that might be viewed as "mechanistic," should require separate PECO statements; however, application of systematic review methods to other mechanistic endpoints, such as mutagenicity, may depend on the needs of the assessment. The key mechanistic and TK questions should be identified to the extent possible in the IRIS assessment plan and IRIS assessment protocol documents.

> **Recommendation 3.10:** When available, KCs should be used to search for and organize mechanistic data, identify data gaps, and evaluate biological plausibility. Those uses should be reflected in the IRIS assessment plan and IRIS assessment protocol.

STUDY EVALUATION

Chapter 6 of the handbook describes approaches for evaluating individual human and animal health effect studies, pharmacokinetic models, and an approach for mechanistic studies. The chapter describes circumstances under which a study may be excluded from the systematic review based on the outcome of the study evaluation. However, such exclusion is inconsistent with recent recommendations to incorporate study evaluation ratings within the context of evidence synthesis.

> **Recommendation 4.1:** The handbook should not use the results of study evaluation as eligibility criteria for the systematic review.

The handbook discusses the term "sensitivity" as the ability of a study to detect a true association, with insensitive studies being prone to producing false negative results. The committee finds the handbook's definition of "sensitivity" to be ambiguous and potentially overlapping with more established systematic review concepts of internal validity, external validity, and statistical precision.

> **Recommendation 4.2:** EPA should evaluate whether aspects currently captured in the notion of "sensitivity" might be better described in the handbook with more established terminology (e.g., precision or generalizability) or better addressed at other points of the systematic review (e.g., risk of bias assessment or evaluation relative to PECO statement[s]). Otherwise, the handbook should provide a more concrete definition of "sensitivity" and a procedure for operationalizing its use in the study evaluation step.

The use of reporting quality as a distinct quality assessment item for study evaluation is not standard for systematic reviews, and procedures for evaluating reporting quality are very different for human epidemiological and animal toxicological studies.

> **Recommendation 4.3:** The handbook should address the apparent difference in assessing reporting quality between the human epidemiological studies and animal toxicological studies by either (1) assessing reporting quality similarly in both types of studies or (2) providing an explicit rationale for why the concepts require different assessment procedures in different types of studies. In either case, the handbook should provide an explicit rationale for isolating elements of reporting quality from established systematic review concepts and evaluate whether aspects currently described as reporting quality might be better addressed at other points of the systematic review process.

EVIDENCE SYNTHESIS

Because many of the considerations for evidence synthesis within a group of outcomes of human or animal evidence are repeated (with slight variation) in Chapters 9 and 11 of the handbook, the transition from the synthesis step to integration is confusing. In addition, the procedures in Chapter 9 do not lead to a strength of evidence judgment for synthesis, which is covered in Chapter 11.

> **Recommendation 5.1:** The handbook should consolidate its discussion of evidence synthesis in a single place. The discussion should include all of the considerations involved in making strength of evidence conclusions (currently in Chapter 9), as well as the criteria for different strength of evidence judgments (currently in Chapter 11). The handbook chapter describing synthesis of evidence should end with the methods for reaching strength of evidence conclusions for each unit of analysis, and how these are carried forward to evidence integration.

The unit of analysis for evidence synthesis and the strength of evidence conclusion is unclear, with respect to the breadth or narrowness of the evidence being synthesized. Although the handbook states that the evaluation of strength of evidence "will preferably occur at the most specific health outcome possible" (EPA, 2020a, p. 11-8), it is not clear how to proceed when there is more than one unit of analysis.

> **Recommendation 5.2:** The unit of analysis for evidence synthesis and strength of evidence conclusions should be clearly defined as specified by the refined PECO statements recommended in Chapter 3 of this report. For example, a unit of analysis could be defined at the endpoint level (e.g., clinical chemistry) or outcome level (e.g., liver toxicity). If judgments may be made at both the endpoint and health outcome levels, details should be provided on how these judgments and the methods used to make them are distinct from each other.

The evidence synthesis approach outlined in the handbook appears to be a hybrid of a guided expert judgment approach and a more structured approach (see Chapter 5 of this report).

Recommendation 5.3: The handbook should provide justification for the initial rating for strength of evidence, as well as more detailed operationalization of the criteria used to upgrade or downgrade the evidence.

The handbook's considerations of mechanistic and TK data, and PK or physiologically based pharmacokinetic (PBPK) models outlined in evidence synthesis appear to mix in some elements of evidence integration, particularly through the concepts of "coherence" of study findings across different endpoints and "biological plausibility" of the findings. The result is confusing, especially when combined with the unclear transition between Chapters 9 and 11, as described above. The application of the term "coherence," as described in the handbook, appears to be more appropriate during either (1) planning of the assessment (the biological relationship among different endpoints) or (2) evidence integration (through incorporation of mechanistic data). Overall, the applications of the term "biological plausibility" in the handbook appear to either (1) address considerations already covered elsewhere, such as consistency, or (2) involve comparison with data on mechanistic changes, which may be more appropriate to incorporate during the evidence integration step.

Recommendation 5.4: The handbook should restrict the applications of mechanistic and TK data or PK models in evidence synthesis and strength of evidence judgments to those relevant to each individual unit of analysis, such as addressing consistency and indirectness of evidence. Other applications of mechanistic and TK data or PK models, such as addressing coherence and elements of biological plausibility, could be addressed in evidence integration, either as a separate evidence stream or as support for the human or animal evidence streams. This recommendation should be implemented in the planning stage and reflected in the protocol (see Recommendation 3.9).

EVIDENCE INTEGRATION

Chapter 11 of the handbook contains information on three sequential but independent steps in a typical systematic review process: synthesizing the evidence (unit[s] of analysis within human and animal streams), rating the confidence in the body of the evidence, and integrating the evidence (integrating human and animal evidence and considering mechanistic evidence). As noted above, there is overlap with handbook Chapter 9 on evidence synthesis. The terms "synthesis," "integration," and "strength of the evidence" appear to be used almost interchangeably throughout these two chapters.

Recommendation 6.1: The handbook should separate and delineate the chapters with respect to evidence synthesis within a stream and evidence integration across data streams (see Recommendation 5.1). Synthesis, integration, and the judgments that are used to rate the evidence need to be clearly defined, distinct steps.

Recommendation 6.2: EPA should supplement Chapter 11 of the handbook with additional figures and examples from IRIS assessment documents. The addition of a terminology map (endpoint, outcome, synthesis, integration) would help define the steps that should occur at each level and, more specifically, what data are to be synthesized and where to expect the judgment narratives to be provided.

Mechanistic data appear to be used to strengthen conclusions of the individual data streams or when there is a lack of evidence in a single stream, and yet some aspects of the handbook treat mechanistic data as their own unit of analysis for evidence synthesis.

> **Recommendation 6.3:** The handbook should clearly define the roles of mechanistic data and other supporting data in evidence integration and throughout the entire IRIS assessment development process (see Recommendations 3.9, 3.10, and 5.4).

STUDY SELECTION FOR DERIVING TOXICITY VALUES AND DERIVATION OF TOXICITY VALUES

EPA has made considerable progress in acting on the 2014 National Academies report recommendation that the IRIS program "continue its shift toward the use of multiple studies rather than single studies for dose–response assessment" (NRC, 2014, p.129) by developing formal methods for combining data from multiple studies. EPA also has developed a process for characterizing uncertainty and communicating confidence in derived toxicity values, as recommended in the 2014 report.

Despite these improvements, the methods described in Chapter 12 of the handbook lack a transparent discussion of considerations used to select studies. Chapter 7 of the committee's report provides some key questions that could be considered to clarify study selection.

> **Recommendation 7.1:** The handbook should provide a clearer, step-by-step description of study selection, using a framework incorporating the different steps of hazard identification (including study evaluation, synthesis, and integration) as well as new steps specific to toxicity value derivation. The handbook should provide a template for how IRIS assessments are to summarize (e.g., in a table) the study selection process as applied to each endpoint, health outcome, study, and evidence stream in order to provide transparency as to study evaluation for toxicity value derivation, and to support selection of overall toxicity values. It is especially important to capture study attributes for which EPA has designated an option as "preferred" versus "less preferred."

Chapter 13 of the handbook provides important information relating to issues and considerations for developing points of departure (PODs) for toxicity values, but it lacks a consistent level of detail for deriving and utilizing PODs.

> **Recommendation 7.2:** EPA should streamline Chapter 13 of the handbook, especially Section 13.2, to focus on the most common methods and approaches rather than detailing less common scenarios. For instance, although use of PBPK modeling is designated as "preferred," it requires too much detail in this handbook to provide instructions on development and application of such models; citing other documents is preferable. If there is important information that is missing from existing EPA documents or from the peer-reviewed literature, these could be provided in an appendix to avoid disrupting the flow of the handbook. Additionally, there may be concerns over providing duplicated information, as any future updates to the related EPA documents would require an update to the handbook as well.

Chapter 13 of the handbook is unclear as to the use of probabilistic approaches to replace the traditional deterministic uncertainty factor-based approach for toxicity value derivation.

> **Recommendation 7.3:** EPA should make it explicit in the handbook that probabilistic approaches to derive risk-specific doses will be routinely applied where feasible, referencing recent literature including a 2020 case study on acrolein (Blessinger et al., 2020). EPA should also consider when and how to transition fully away from the traditional deterministic approach to adopt risk-specific doses for its IRIS toxicity values.

CONCLUDING REMARKS

Overall, the committee concluded that the handbook reflects the significant improvements that EPA has made in its IRIS assessment process. The methods for developing IRIS assessments can serve as a model for other EPA programs that are implementing systematic review methods. The committee believes that the recommendations provided in this report will help ensure that the handbook meets its objectives of providing transparency about the IRIS assessment process and providing operational instructions for those conducting the assessments.

1

Introduction

The U.S. Environmental Protection Agency's (EPA) Integrated Risk Information System (IRIS) Program is the part of the agency's Office of Research and Development (ORD) that develop human health assessments that focus on hazard identification and dose-response analyses for chemicals in the environment. IRIS assessments cover cancer and noncancer outcomes and include toxicity values (reference values and unit risks) that can be combined with exposure estimates to develop quantitative estimates of risk. The assessments are highly important as they are used to inform risk assessments and risk management decisions throughout the agency, including EPA regulatory programs. The assessments are also used by federal, state and tribal agencies as well as community organizations and agencies in other countries to inform decisions concerning health risk assessment and management.[1]

Over the years, questions had been raised about the scientific basis of toxicity values reported in some IRIS assessments and the extensive amount of time taken to complete assessments. A 2011 report from the National Academies of Sciences, Engineering, and Medicine that reviewed a draft IRIS assessment of formaldehyde offered recommendations for improving the IRIS process (NRC, 2011). In 2014, the National Academies issued a report that reviewed the IRIS program process and expanded on those recommendations. Although many other recommendations were provided in the 2014 report, it concluded that "substantial improvements in the IRIS process have been made, and it is clear that EPA has embraced and is acting on the [National Academies] recommendations" (NRC, 2014, p. 9). In 2018, a National Academies report reviewed progress toward transforming the IRIS program and concluded that substantive progress has been made in transforming the IRIS program to date. It also found the program to be on track toward full implementation of systematic review and transparency in IRIS assessments (NASEM, 2018).

A major recommendation of the 2014 National Academies report was to adopt systematic review methods throughout the IRIS assessment process, including the steps on scoping, problem formulation, protocol development, study identification and evaluation, and evidence integration (NRC, 2014). Systematic review is "a scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies. The goal of systematic review methods is to ensure that the review is complete, unbiased, reproducible, and transparent" (IOM, 2011, p. 1).

The 2014 National Academies report recommended that EPA develop a handbook to "provide a single detailed guidance document for all those involved in the development of IRIS assessments" and make the IRIS process transparent to stakeholders (NRC, 2014, p. 23). The 2014 report acknowledged the challenges of having one guidance document to meet the needs of EPA staff and consultants who conduct IRIS assessments, as well as stakeholders.

A 2021 National Academies report that reviewed EPA's Toxic Substances Control Act (TSCA) Risk Evaluations (NASEM, 2021) also recommended the development of a handbook

[1] Additional information on the IRIS program is available at https://www.epa.gov/iris.

for EPA's TSCA review and evidence integration methodology. TSCA risk evaluations are conducted by EPA's Office of Pollution Prevention and Toxics (OPPT), which is separate from the IRIS Program. The 2021 report also recommended that OPPT review its approach and incorporate components of the National Institute of Environmental Health Science's Office of Health Assessment and Translation systematic review and evidence integration method (NTP, 2019), the University of California San Francisco's Navigation Guide (Woodruff and Sutton, 2014), and IRIS methods.

The *ORD Staff Handbook for Developing IRIS Assessments* (the handbook), issued in November 2020, provides guidance to scientists who perform the IRIS assessments in order to foster consistency in the assessments and enhance transparency about the IRIS assessment process (EPA, 2020a; Thayer, 2021). The handbook does not supersede existing EPA guidance and does not serve as guidance for other EPA programs. It also provides an opportunity for stakeholder communities to become aware of the processes and policies guiding those who are drafting IRIS assessments.

THE COMMITTEE, ITS TASK, AND ITS APPROACH

EPA requested that the National Academies conduct a review of the handbook (EPA, 2020a). In response, the Committee to Review EPA's IRIS Assessment Handbook was convened to evaluate the operating procedures, as described in the handbook, for hazard identification and dose-response assessment of potential human health effects from exposure to environmental contaminants. The committee was also asked to review the procedures and considerations for operationalizing the principles of systematic reviews, and the methods described in the handbook for determining the scope of the IRIS assessments (e.g., health outcomes, routes of exposure), evidence evaluation, evidence integration, and dose-response analyses. The committee's verbatim statement of task is provided in Box 1-1. The committee included experts in epidemiology, pharmacology, toxicology, statistics, environmental health, dose-response modeling, pharmacokinetic and pharmacodynamic modeling, systematic review, and risk assessment. (See Appendix A for biographical information on the committee.)

EPA provided the committee with a list of detailed questions to further guide the committee's review that, with one exception (see Chapter 6 of this report), fall within the general outline of the statement of task. Appendix B provides brief descriptions of the organization and content of the handbook chapters and a listing of EPA's questions for the committee.

BOX 1-1 Statement of Task

An ad hoc committee of the National Academies of Sciences, Engineering, and Medicine will evaluate EPA's ORD Staff Handbook for Developing IRIS Assessments (or IRIS Handbook). The committee will evaluate the operating procedures for developing hazard and dose-response assessments of potential human health effects from exposure to environmental contaminants. The procedures and considerations for operationalizing the principles of systematic review within these human health assessments will be evaluated. Consideration will also be given to the guidance on determining the scope of the IRIS assessments (e.g., health outcomes, routes of exposure), evidence integration; extrapolation techniques (e.g., physiologically based pharmacokinetic (PBPK) models), dose-response analyses (e.g., statistical methods), and characterization of uncertainties.

In carrying out its task, the committee held eight meetings, including information-gathering sessions on February 11, 2021, and April 16, 2021, to hear presentations from Jonathan Samet, Colorado School of Public Health, and Kris Thayer, EPA, respectively. The committee also heard public comments during each meeting (see Appendix C for meeting agendas). EPA's presentation addressed questions from the committee regarding the handbook (Thayer, 2021). In addition, the committee received various written materials, including EPA's written responses to the committee's questions and the public comments on the handbook that were submitted to the EPA docket.[2]

The committee reviewed the handbook, recognizing that the primary audience is internal to EPA but also includes external stakeholders. It developed a list of considerations (Box 1-2) to guide its critique of each chapter of the handbook. The list includes transparency; clarity; organization; terminology; areas of controversy about the method; and whether the described method was accurate, complete, and could be applied by those responsible for conducting IRIS assessments. Not all considerations were relevant to each handbook chapter.

BOX 1-2 Considerations for Critique of Handbook Chapters

1. Given the committee's understanding of the handbook, is the method (process) clearly described? Are the objectives of the chapter clear?

2. Will following the handbook methods result in transparent, credible reviews of the evidence that are of high utility given the remit of the IRIS program?

3. Are the described methods aimed at the correct objectives?

4. Is the terminology used in a precise, consistent manner?

5. Are the described methods likely to be implemented in a reasonably efficient manner?

6. Do the methods align with state-of-the-science/valid/reliable methods in the general field? Are they sufficiently rigorous?

7. Are key aspects of the methods missing or interpreted incorrectly in the handbook?

8. Do any choices of methods warrant further justification? Are any EPA policy decisions that restrict the choice of method adequately explained in the handbook?

9. Are any aspects particularly controversial (e.g., ongoing debate about which method among several options is optimal)?

10. Are the steps involved in implementing a method described at an appropriately granular level?

11. Is the method description specific enough such that the method can be carried out in a repeatable, transparent manner by a handbook user?

12. Does the method allow for some flexibility in application for a wide variety of IRIS assessments (e.g., variability in the types and amounts of evidence for an assessment)?

13. Are there methodological developments that EPA should be aware of, as they continue to revise the handbook? Distinguish between methods that can be applied now and those that might be available in the future.

[2] A 90-day public comment period was associated with the release of the handbook. See Federal Register vol. 85, No. 230, Nov 30, 2020. 76566–76567.

The committee's review of the handbook is independent of the 2021 TSCA review, although the committee's findings and recommendations could help inform OPPT's responses to recommendations in NASEM (2021) and harmonize EPA's approach to hazard evaluation.

ORGANIZATION OF THE REPORT

The committee structured its report chapters according to the major steps in the IRIS process for developing assessments. Chapter 2 focuses on an overview of the handbook's organization and content. Chapter 3 considers the handbook's coverage of planning IRIS assessments, including evidence mapping, scoping, protocol development, the refined evaluation plan, and organization of the IRIS assessment. Chapter 4 focuses on the handbook's treatment of evaluation of individual studies within a data stream. Chapter 5 covers the handbook's handling of evidence synthesis and judgments about the strength or certainty of the synthesized evidence within a data stream. Chapter 6 focuses on the handbook's treatment of evidence integration across data streams and the judgments used to describe the strength of conclusions regarding hazard identification. Chapter 7 covers the handbook's steps of selecting studies for toxicity derivation and deriving toxicity values based upon dose-response analyses. Table 1-1 maps the chapters of the committee's report with the questions from EPA and relevant handbook chapters.

As requested by EPA (see Appendix B), the committee organized its recommendations to indicate their relative importance during EPA's revision of the handbook:

- *Tier 1: Recommended Revisions*—Highest priority recommendations the committee believes are critical to improve the scientific rigor and/or clarity of the document.
- *Tier 2: Suggestions*—Recommendations that EPA should consider to strengthen the document.
- *Tier 3: Future Considerations*—Topic areas that may inform future developments.

These recommendations are outside the immediate scope and/or needs of the current document under review.

TABLE 1-1 Mapping of Committee's Report Chapters and EPA Questions

Committee Report Chapter	Questions from EPA[a]	Relevant Handbook Chapter
2. Overview of Organization and Content of the Handbook	Q1. Comment on the overall organization of the handbook, in particular on whether key aspects of the assessment process are represented.	Entire Document
3. Planning Assessments	Q2. Are systematic review approaches used by the IRIS Program (outlined in Chs. 1-5) clearly described and consistent with methodologies considered to be state-of-the-science by experts in the field? Q4. Does the handbook clearly lay out a state-of-the-science approach for refinement of the scope and analyses of the IRIS assessment? Are there specific areas for improvement (recommended alternatives)? Q5. Review and assess the handbook's process for evaluating and integrating mechanistic data. [Sections 2.2, 4.3.3, and 6.6, and Ch. 10] Are there specific areas for improvement (recommended alternatives)?	1-7, 10
4. Study Evaluation	Q3 (part b). Are the study evaluation methods in Ch. 6 for individual human studies (epidemiology and controlled exposure) and animal studies mechanistic evidence (pilot testing approaches), and pharmacokinetic models adequate? If not, how can the methods be improved?	6
5. Evidence Synthesis	Q6 (part a). Are the approaches to evidence synthesis described in Ch. 9 and Ch. 10 scientifically sound? • Are the considerations sufficiently broad to allow for application to the wide range of scenarios that will be encountered when applied to individual assessments? • Are the methods sufficiently clear in describing the intent to synthesize the relevant evidence, incorporating study evaluation conclusions, regardless of the study results? [NOTE: Ch. 11 discusses synthesis as part of integration]	9-11
6. Evidence Integration	Q6 (part b). Are the approaches described in Ch. 11 scientifically sound and appropriate for integrating the various types of evidence relevant to investigating the potential for human health effects from exposure to environmental chemicals? Q7. Comment on approaches using five categories versus three categories for drawing evidence integration conclusions. Which approach is recommended and why? Are there any specific refinements for improvement?	11

7. Hazard Considerations and Study Selection for Deriving Toxicity Values	Q8. Does the handbook provide appropriate considerations in Ch. 12 and Ch. 13 for identifying data sets for dose-response analysis based on systematic review conclusions? Are the basic methods for dose-response modeling and deriving toxicity values consistent with the current state of the science, and presented with sufficient clarity?	12, 13

[a] Chapter and section numbers listed in the questions refer to those of the *ORD Staff Handbook for Developing IRIS Assessments* (EPA, 2020a). See Appendix B of this report for additional information.

2

Overview of Organization and Content of the Handbook

This chapter evaluates the organization and overall content of the *ORD Staff Handbook for Developing IRIS Assessments* (the handbook) (EPA, 2020a) and whether key aspects of the assessment process are represented (Question 1 in Table 1-1). In doing so, the chapter specifically considers the Integrated Risk Information System (IRIS) assessment process illustrated in Figures O-1 (p. xviii) and O-2 and Table O-1(p. xix) in the handbook.

Overall, the committee found that the handbook reflects the significant improvements the U.S. Environmental Protection Agency (EPA) has made in its IRIS process for developing draft assessments. For instance, the process includes sophisticated, state-of-the-art methods related to using systematic evidence maps for scoping and systematic review methods for hazard identification. Moreover, the IRIS program is clearly helping to advance the science of systematic review as applied to hazard identification. EPA staff are actively involved in the ongoing development of methods, such as study evaluation and handling of mechanistic data. The committee recognizes that implementation of many of the methods now used in the IRIS assessment process is challenging for EPA, as they have not been previously used by the agency and some methods are still evolving. The committee is impressed and encouraged by the progress the IRIS program has made to date. The IRIS process for developing assessments can serve as a model for other parts of EPA that are implementing systematic review methods.

ORGANIZATION OF THE HANDBOOK

The strengths and advances in methodology for the IRIS assessment process are unevenly and not always clearly conveyed in the handbook. The handbook alone does not adequately describe the overall flow of the process for developing an IRIS assessment. Thus, more could be done to ensure that the handbook meets its objectives of providing transparency about the process and fostering consistency in assessments developed by the IRIS program by operationalizing the process. The presentation (Thayer, 2021) and written materials that EPA provided to the committee were important resources that enabled it to understand some sections of the handbook. Inclusion of information and examples from Dr. Thayer's presentation into the handbook would improve its clarity. For example, the presentation used screenshots from the Health Assessment Workspace Collaborative (HAWC) software used by EPA for IRIS assessments to illustrate how study evaluations within and across evidence stream judgments were carried through the IRIS process. The committee understands that HAWC is a flexible program that can be adapted to accommodate changes in methods. Thus, it is important that illustrations from HAWC match the guidance in the handbook.

The committee also acknowledges that the IRIS assessment plan (IAP) published for each IRIS assessment provides details on specific methods and approaches that are employed for specific IRIS assessments. Nevertheless, if each step in the IRIS assessment process were clearly outlined in the handbook, it would make the IRIS assessment process more transparent while

simultaneously benefitting the handbook users who conduct IRIS assessments. The handbook chapters for specific steps in the IRIS assessment process lack a general description of the methods used for the step and a clear outcome expected at the completion of the step. For example, handbook Chapter 8, "Data Extraction," describes the process for extracting data from studies that have been identified as meeting PECO (population, exposure, comparator, outcome) inclusion criteria and being "sufficiently informative," as described in Chapters 4 and 6 of the previous handbook. Handbook Chapter 8 also addresses methods for standardizing the presentation of effect sizes and doses, and data display options. These processes are heavily dependent on the use of HAWC. The level of detail presented in Chapter 8 obscures how and where data extraction fits into the overall assessment process. If much of the detail about HAWC and data display were presented as supplementary material, the chapter could focus on the conceptual content related to standardizing effect sizes and doses.

The committee also found that the poor organization of parts of the handbook text detracts from its readability. For example, numerous "call-backs" and "call-forwards" across multiple chapters make the handbook difficult to navigate and the process less transparent. Moreover, differences in the level of detail, a degree of repetition, and occasional inconsistencies in information across some chapters also make the handbook difficult to follow.

Chapter 7 of the handbook, "Organizing the Hazard Review: Approach to Synthesis of Evidence," appears to be discussed as part of the planning process where outcomes are prioritized for synthesis, analysis plans are developed, and next steps are refined and prioritized. The outcome of that chapter and how it feeds into the revision of the IRIS assessment protocol are not clearly presented in the handbook.

Another example of a need for improved organization of the handbook is the overlap between Chapter 9, "Analysis and Synthesis of Human and Experimental Animal Data," and Chapter 11, "Evidence Integration." Both chapters list considerations for evidence synthesis within a group of outcomes of human or animal evidence (e.g., Table 9-1 (p. 9-3) and Table 11-2 (p. 11-10)), but the criteria are not the same in each chapter. Only Chapter 11 describes how the ratings for each of these considerations produces a strength of evidence judgment for a body of human or animal evidence, which is then advanced to the evidence integration step.

ILLUSTRATIONS OF THE IRIS ASSESSMENT PROCESS

Figures O-1 and O-2 and Table O-1 in the handbook portray the process for developing a draft IRIS assessment as strictly linear. However, as mentioned in the handbook text, the process is iterative. Searches for studies and evidence mapping can identify a wider range of studies than those eventually included in evidence synthesis and integration. Some studies, which may be considered supplemental at the evidence mapping stage, may be included in the IRIS assessment at a later stage. For example, additional studies that provide information on mechanisms or toxicokinetics (TK)[1] may enter the IRIS process at a later stage. The handbook does not provide a clear illustration of how the number of studies initially identified is narrowed or expanded throughout the assessment process. In addition, the handbook does not clearly illustrate when and where in the process the outcomes for the PECO questions for the systematic reviews for

[1] In this report toxicokinetics refers to the absorption, distribution, metabolism, and excretion processes also called ADME or pharmacokinetics. However, consistent with the handbook, models are described as pharmacokinetic (PK) or physiologically based pharmacokinetic (PBPK) models rather than toxicokinetic models.

hazard identification are defined or refined. The systematic review protocol is shown as a single step in the IRIS assessment process, but multiple systematic review protocols could be part of a single IRIS assessment. EPA noted that the IAP is a research protocol for scoping and problem formulation, where systematic evidence maps are used increasingly (Thayer, 2021). Producing evidence maps can enhance transparency regarding the available data, but the use of evidence maps is not well described in the handbook. In addition, the handbook does not adequately indicate how the units of analysis are eventually selected for hazard identification and toxicity value derivation.

EPA explained that the IRIS assessment protocol has initial and final versions that describe the systematic review and dose-response methodologies (Thayer, 2021). Both versions of the protocol include PECO, literature searching and screening, literature inventory to describe included studies, tagging of supplemental materials, study evaluation, data extraction, and evidence synthesis/integration. The final version of the IRIS assessment protocol may encompass revisions made in consideration of public comments received on the initial protocol. The final version of the protocol also provides more specific details on methods that could not be fully described during the initial version, such as analysis of mechanistic and TK data and non-standard dose-response methodology. The final version could also include the list of prioritized outcomes/endpoints, but as noted elsewhere in this report the process for prioritizing the endpoints may not be clear.

OVERALL CONTENT OF THE HANDBOOK

The committee found that the handbook does not focus sufficiently on the major steps in the IRIS assessment process: planning the assessment, study evaluation, evidence synthesis, evidence integration, and selecting and deriving toxicity values . Streamlining the handbook to concentrate on describing the key concepts, definitions, and instructions required to complete those major components of the IRIS process would improve the transparency and usability of the handbook. A professional editor could assist with this task and advise on whether details on methods that are rapidly evolving, such as evaluation of human epidemiology studies of environmental exposures or mechanistic study evaluation, would be better placed in supplementary materials, where they could be updated regularly, rather than in the main text.

The clarity of the handbook was also hindered by the lack of a glossary of key terms. Definitions for key terms such as "scoping," "synthesis," "integration," and "sensitivity" are currently scattered across different chapters. In addition, the handbook sometimes uses terminology in an inconsistent and unconventional manner, with a tendency to use terms with pre-established definitions in the general area of evidence synthesis to designate various IRIS-specific processes and products. Examples include the use of the term "protocol" to describe the general plan for an IRIS assessment rather than a detailed plan for conducting a specific systematic review; the use of the terms "preliminary literature survey," "systematic evidence map," and "literature inventory" to refer to an evidence database, when all three terms have different established meanings; and the use of "scoping review" to refer to a complex process that includes engaging stakeholders in the identification of programmatic needs, rather than a survey of a body of evidence conducted to inform a research planning process.

Use of Systematic Review Methods

As recommended in previous National Academies reports, systematic review methods are being incorporated into the IRIS assessment process. These reports have defined systematic review as a "scientific investigation that focuses on a specific question and uses explicit, prespecified scientific methods to identify, select, assess, and summarize the findings of similar but separate studies" (IOM, 2011, p. 1). In practice, the IRIS program is applying systematic review methods appropriately to parts of the assessment process (e.g., evidence evaluation and synthesis for hazard identification). However, in some places, the handbook suggests that systematic review methods are being used at all points in the assessment process, which is not actually the case.

In particular, Figure O-1 suggests that systematic review methods are used for all steps from developing the literature identification strategy through selection of studies for dose-response assessment. However, not all of these steps in the IRIS process are amenable to systematic review methods. For example, although evidence mapping is a rigorous method used for scoping, it is not a systematic review method. As noted in the handbook, systematic review methods are not directly applicable for conducting dose-response analyses, although studies that form the basis of toxicity value derivations should have gone through a systematic review process during hazard identification. Indeed, all studies that may be used for toxicity values , whether human, animal, or mechanistic, are expected to go through a systematic review process before determination of hazard identification. On the other hand, some mechanistic and TK studies that enter the IRIS process provide supplementary information (e.g., on biological plausibility) and may or may not need to go through a systematic review process, depending on the needs of the assessment. The handbook would be more useful and transparent if it clearly delineated where systematic review methods are appropriate and how they are applied in the IRIS process.

Mechanistic and Toxicokinetic Data

Multiple places in the handbook describe the roles of mechanistic data, TK data, and results from PK models and physiologically based pharmacokinetic (PBPK) models. However, the descriptions are not entirely consistent. For instance, Table 2-2 of the handbook (p. 2-7) provides examples of TK studies and PK and PBPK studies, but this information is repeated with some variation in Section 4.2.1. The main discussion of mechanistic information is deferred until Chapter 10 of the handbook, leading to multiple instances of both backward and forward referencing, which makes the handbook confusing as to when and how mechanistic and TK data are incorporated into the assessment. The readability of the handbook would be improved by specifying, near the beginning of the handbook, the range of potential roles for mechanistic, TK, and other related data in the assessment (e.g., see Table 2-1 of this report), and then carrying that information consistently throughout the rest of the document.

TABLE 2-1 Common Uses of Mechanistic and Toxicokinetic Evidence in IRIS Assessments

Mechanistic Evidence	Toxicokinetic (ADME) Evidence[a]
Hazard Evidence Synthesis: • Identity of active moiety of an agent • Identity of target tissue • Key characteristics • Interspecies differences (including human relevance) • Mode of Action/Adverse Outcome Pathway analysis • Endpoints that may be linked/related/grouped together • Potential for read-across for predicting certain endpoints, based upon well-studied similar chemicals • Susceptible populations and life stages	**Hazard Evidence Synthesis:** • Active moiety • Target tissue concentration • Route of exposure differences • Interspecies differences • Susceptible populations and life stages
Dose-Response: • Low-dose extrapolation approach • Interspecies differences in sensitivity • Dose-metrics for each endpoint • Susceptible populations and life stages	**Dose-Response:** • PK or PBPK model(s) to predict dose metrics • Route of exposure differences/extrapolation • Interspecies extrapolation • Susceptible populations and life stages

[a] Includes information relating to PK and PBPK models.

The handbook does not adequately describe the role of mechanistic data throughout the IRIS assessment process. For example, it is not always clear when mechanistic data would be treated as a separate data stream for hazard identification, when it would be used to increase or decrease confidence in animal or human studies based on biological plausibility, and when it would potentially serve both roles. Despite this shortcoming, it is laudable and significant that the IRIS program is striving to incorporate mechanistic data in a manner that is consistent with efforts to move toward decreased reliance on animal-based studies and increased use of New Approach Methods (NAMs), such as those that rely on in silico or in vitro methods. An increased use of mechanistic data in the IRIS assessment process could encourage its use more broadly and serve as a model for other agencies conducting hazard evaluations.

Susceptible Populations

Most risk assessments involving environmental, industrial, or occupational chemical exposures have the potential to involve susceptible life stages or other types of susceptibility (e.g., genetic polymorphisms; aging systems and pre-existing disease; and social, lifestyle, and demographic factors [as described in the handbook's Table 9-2 on p. 9-6]). Identifying susceptible populations and determining whether there are sufficient data to evaluate dose-response relationships in such populations are challenges common to most, if not all, chemical assessments. Those challenges are particularly important for susceptible life stages since the period of susceptibility may be relatively brief and thus may not involve consideration of long-term exposure. This has implications for the type of age-dependent adjustment factors that may be needed for cancer assessments and the shorter-term exposure parameters that may be needed to align with the susceptibility window. The extent, nature, quality, outstanding issues, and uncertainties with respect to the evidence base for susceptible populations are important considerations.

The handbook recognizes the importance of assessing potentially susceptible populations by encouraging their consideration throughout the assessment process, and the handbook describes in multiple places how the evidence base for susceptible populations should be handled. However, the handbook treats this aspect of hazard identification as a special case evaluation that may not be required unless animal, human, or mechanistic evidence points to a particular susceptibility or life stage. Discussions that define what constitutes evidence of susceptibility and describe the types of data that may inform such susceptibility are not presented in one place in the handbook, but those topics are alluded to in various places throughout the document. For example, the handbook describes refinements to the evaluation plan (handbook Chapter 5) to include "studies that address critical lifestage or exposure duration-specific knowledge of the development of the health outcome (e.g., for endpoints relating to organ development or cancer, respectively)" (EPA, 2020a, p. 5-2). In Chapter 9 of the handbook (Table 9-2, p.9-6), several other factors are described that may contribute to susceptibility. While it is helpful to alert handbook users to such factors, the handbook does not provide a process for formally considering the susceptibility factors that apply to a given assessment, data available to inform such a consideration, and the needed refinements to PECO statements and the assessment plan.

Linkage of Dose-Response with Hazard Identification

The handbook chapters on selection of studies for toxicity assessment (Chapter 12) and derivation of toxicity values (Chapter 13) are less well developed than the other chapters devoted to hazard identification. Handbook Chapter 12 on selection of studies is not clearly linked to the earlier handbook chapters. For example, it is unclear how the results of the systematic reviews that are conducted for hazard identification are used to select studies for the dose-response assessment. The transition from hazard identification to derivation of toxicity values needs to be improved, particularly with respect to selecting studies for dose-response. By contrast, much of the material in Chapter 13 on deriving toxicity values is described in other EPA documents but not the handbook. Thus, the description of the process for deriving toxicity values could be streamlined substantially to focus on the most common practices, while noting by reference where less common approaches based on mechanistic, TK, or novel quantitative methods may be obtained and applied.

The committee concluded that the transparency and usability of the handbook could be improved by streamlining all of its text to focus on the major steps in the IRIS process, eliminating repetition among the chapters, incorporating examples from IRIS assessments and software tools such as HAWC, and ensuring that terminology is used consistently across chapters.

GAPS IN THE HANDBOOK

The committee identified several key gaps in the content of the handbook: handling of publication bias and funding bias, planning for handbook updates, preparing for new or emerging types of studies that may be used in future IRIS assessments, and including quality assurance in the IRIS assessment process.

Publication Bias and Funding Bias

Publication bias is the publication or non-publication of research results based on the nature and direction of the results (Higgins and Thomas, 2019). Funding bias refers to an association between study funding sources and financial ties of investigators with research outcomes that are favorable for the sponsors (Holman et al., 2019). NRC (2014) recommended that publication bias and funding bias be addressed in IRIS assessments. Handbook Section 8.1.1 notes that the data collection tool, HAWC, allows for collection of information on study funding source and author conflicts of interest, but this does not appear to be mandatory. Author conflicts of interest are mentioned in Section 9.4.3 as a possible source of reporting or publication bias, but funding bias is not mentioned at all. Publication bias is mentioned in the handbook chapters on Evidence Synthesis (Chapter 9) and Evidence Integration (Chapter 11), particularly in relation to how it could impact the assessment of consistency across studies. Funding bias is not mentioned at all in the handbook.

The NRC (2014) recommendation to address funding bias as part of the study evaluation process was based largely on evidence obtained from human studies because less was known about the extent of funding bias in animal research. Since the 2014 report was published, evidence for both publication and funding bias in human and animal literature has increased. Meta-research studies continue to identify publication bias (Rezende et al., 2018; van der Naald et al., 2020) and funding bias (Huss et al., 2007; Anglemyer et al., 2015; Bero et al., 2015; Chartres et al., 2016; Friedman. and Friedman, 2016; Lundh et al., 2017; Wells, 2017) across a variety of fields. Other methodologies, such as case studies, qualitative analysis of interview data from researchers, or internal industry documents, have described specific studies that have been influenced by industry or not published (for example, see Michaels (2006)). There remains a need to conduct meta-research evaluations for some types of studies included in IRIS assessments. For example, a major source of data for chemical assessments conducted by IRIS is standardized animal toxicity studies, typically done by contract research organizations (CROs) under good laboratory practices. The same CROs may be funded by various sources, including industry and government, often simultaneously, and evidence for the influence of funding source in such cases is lacking. Funding bias is important to consider regardless of the funding source.

Methods to detect and evaluate impacts of publication bias or funding bias are available or are being developed. The handbook (Section 9.4.3) notes methods to detect publication bias and assess its impact on evidence synthesis. However, it does not provide guidance on whether or how to apply these methods to assess the impact of publication bias on IRIS assessments. Although most current study evaluation tools are not adequate for addressing bias related to conflicts of interest (Lundh et al., 2019), some tools are available for incorporating assessments of funding bias or author conflict of interest into study evaluation (SIGN, 2011; Woodruff et al., 2011; Moga et al., 2012; Downes et al., 2016). The Navigation Guide (Woodruff and Sutton, 2014, for example, includes author conflicts of interest and study funding source as components of study evaluation (Woodruff et al., 2011), although studies are not excluded based on a high risk of bias in any single domain. Cochrane is developing a Tool for Addressing Conflicts of Interest in Trials (TACIT)[2] to evaluate the occurrence and impact of funding bias and author conflicts of interest on systematic reviews of trials. Funding and publication biases can also be investigated at the evidence synthesis and integration stages. For example, the impact of funding bias on effect estimates obtained by evidence synthesis can be evaluated by excluding studies by

[2] The TACIT website is https://tacit.one/.

funding source in a sensitivity analysis or by conducting subgroup analyses by funding type. Incorporating assessments of publication bias or funding bias at the evidence evaluation or synthesis stage could influence the confidence in estimated effects.

Updating the Handbook

The preface of the handbook indicates that the handbook is a "living document" and "the IRIS program will update the IRIS handbook as needed for major shifts in approaches based on emerging science and experience gained through its application to a broader spectrum of assessments" (EPA, 2020a, p. xiv). Given the current complexity of the handbook and EPA's existing review process, it is difficult to see how the handbook could be updated in a timely manner. The handbook does not describe a process for updating, which would include proposed timelines, how to identify major changes that would need to go through external peer review, and how to more quickly update details of evolving methods (e.g., study evaluation and software tools) that could be linked to the handbook without the need for external review.

Anticipating Future Types of Studies

A handbook chapter on preparedness for incorporating future types of studies would demonstrate EPA's involvement in the development of emerging methods. Currently, human epidemiology and animal toxicology studies are the main focus of the handbook. A preparedness chapter could identify additional types of studies that EPA is considering for future IRIS assessments (e.g., high throughput, non-rodent vertebrate studies, and in silico). If NAMs were to be recognized as eventual replacements for animal-intensive toxicology studies, emerging methodologies for NAMs of potential future relevance to IRIS could be highlighted. Evaluation of chemical mixtures is a gap in the handbook and may be beyond the current scope of IRIS assessments. The handling of life stage studies, including developmental stage studies, in IRIS assessments is not fully developed in the handbook. The handbook does not recognize new methodological developments in those areas and does not target them for possible future use.

In addition, systematic reviews of human, animal, mechanistic, or other types of relevant data are becoming more common in the general field of environmental health. Incorporating systematic reviews from other sources into IRIS assessments could enhance process efficiency.

Quality Assurance

Previous National Academies reports have noted weaknesses in the quality assurance of the IRIS process. NRC (2014) recommended that EPA provide a quality management plan that included clear methods for continuous evaluations of the quality of the IRIS process. NASEM (2018) found the IRIS management had taken multiple steps to ensure high-quality, including structures to improve the quality of the IRIS assessments.

The handbook can play a key role in the quality assurance of the IRIS assessment process because it can standardize the practices of IRIS program staff and contractors working on IRIS assessments. However, the handbook lacks a section on the overall quality assurance of the assessment process, how quality will be monitored, and how EPA staff and consultants will be trained, as needed, to meet the quality assurance standards.

The handbook includes quality assurance procedures related to some individual steps in the process. For example, handbook Chapter 6 indicates that the quality assurance of the study evaluation process is accomplished by having each study evaluation generally conducted independently by at least two reviewers, with a process for comparing and resolving differences. However, the handbook provides the option to use only one reviewer, which is not standard practice for systematic reviews, and fails to note which of the described methods have been tested empirically. Empirically based methods that are valid and reliable could establish a minimum set of standards for quality assurance. Alternatively, pragmatic standards, developed collaboratively by methodologists and those who conduct IRIS assessments, could be established. For example, Cochrane and others have developed methodological expectations for the conduct, reporting, and updating of systematic reviews (Schaefer and Myers, 2017; Higgins and Thomas, 2019; Whaley et al., 2020). The Cochrane standards are categorized as "mandatory" or "desirable" for systematic review publication.

FINDINGS AND RECOMMENDATIONS

General Finding: The handbook reflects the significant improvements EPA has made in its IRIS assessment process. For instance, the process includes sophisticated, state-of-the-art methods related to using systematic evidence maps for scoping and systematic review methods for hazard identification. Moreover, the IRIS Program is clearly helping to advance the science of systematic review methods as applied to hazard identification. EPA staff are actively involved in the ongoing development of methods, such as study evaluation and use of mechanistic data. The IRIS assessment methods can serve as a model for other parts of EPA that are implementing systematic review methods.

Findings and Tier 1 Recommendations

Finding: The objectives of the handbook are to improve transparency about the IRIS assessment process and provide operationalizable instruction for those conducting IRIS assessments. Achieving both of these objectives in a single document may be difficult.

> **Recommendation 2.1:** EPA should engage a professional editor with specific expertise in developing handbook-like materials to assist with the handbook revision. The editor should enhance the transparency and ease of use of the handbook by focusing the material on the concepts, definitions, and instructions needed to complete the main steps in the IRIS assessment process; eliminating unnecessary repetition among the chapters; and ensuring that terminology is used consistently across chapters. *[Tier 1]*

> **Recommendation 2.2:** EPA should add a glossary to the handbook for defining key terms. Single definitions should be provided for concepts, and the definitions should be applied consistently throughout the handbook. *[Tier 1]*

Finding: The handbook uses unconventional terminology, with a tendency to use terms with pre-established definitions in the area of evidence synthesis to designate various IRIS-specific processes and products. Examples include the use of the term "protocol" to describe the general plan for an IRIS assessment, rather than a detailed plan for conducting a specific systematic

review; the terms "preliminary literature survey," "systematic evidence map," and "literature inventory" to refer to an evidence database, when all three terms have different established meanings; and "scoping review" to refer to a complex process that includes engaging stakeholders in the identification of programmatic needs, rather than a survey of a body of evidence conducted to inform a research planning process.

Systematic review methods are used for parts of the overall IRIS process. However, the handbook suggests, at some points, that systematic review methods are being used when they are not. The IRIS program is applying systematic review methods appropriately to parts of the process (e.g., evidence evaluation and synthesis for hazard identification), However, the handbook does not clearly differentiate when systematic methods (such as a systematic search of the literature or another method that is applied systematically) versus systematic review methods are being used throughout the IRIS process.

> **Recommendation 2.3:** The handbook should use terminology in a manner that is consistent with existing, accepted definitions in related fields.. When alternative definitions are used for the IRIS assessment process, the handbook should provide explicit justification. *[Tier 1]*

> **Recommendation 2.4:** When systematic review methods are being used for parts of the IRIS assessment process, this should be stated and the relevant methodological literature should be referenced. *[Tier 1]*

Finding: The handbook does not adequately describe the overall process or flow for developing IRIS assessments and the iterative nature of some of the steps in the process.

> **Recommendation 2.5:** EPA should create new graphical and tabular depictions of the IRIS assessment process as it is currently practiced, and should not feel constrained to mirror the process depicted by the 2014 National Academies report *Review of EPA's Integrated Risk Information System (IRIS) Process* or any other report (including this one). A professional editor could be of assistance here. *[Tier 1]*

Finding: The handbook includes very detailed information on some methods that may undergo rapid development (e.g., data extraction). On the other hand, the handbook lacks specific examples from relevant IRIS assessments and examples of the software used by EPA, such as HAWC, that are needed to fully understand the IRIS program's methods.

> **Recommendation 2.6:** The handbook should incorporate more examples from relevant IRIS assessments and examples of software used by EPA, such as HAWC. EPA could provide them as supplementary material or links to other content. *[Tier 1]*

Finding: Publication bias and funding bias are mentioned sparingly, or not at all, in the handbook, although empirical evidence from a variety of fields shows that funding bias and publication bias can alter effect estimates when evidence is synthesized. Funding bias might also have an effect on the confidence of study ratings from evidence evaluation. Tools and methods to detect and assess these biases are available.

Recommendation 2.7: The handbook should describe how to detect and assess the effect of funding bias on the confidence of study ratings from evidence evaluation or effect estimates from synthesis. *[Tier 1]*

Recommendation 2.8: The handbook should describe how to detect and assess the effect of publication bias on effect estimates from synthesis. *[Tier 1]*

Findings and Tier 2 Recommendations

Finding: The roles of mechanistic, TK information throughout the IRIS assessment process are not clearly described in the handbook. For instance, the main discussion of the use of mechanistic data is deferred until Chapter 10, so that discussions of those types of data earlier in the handbook require call-forwards and do not have adequate context to be clearly understandable. In addition, it is not clear when mechanistic data could be treated as a separate data stream for hazard identification, could be used to increase or decrease confidence in animal or human studies based on biological plausibility, or could be used to inform key science considerations.

Recommendation 2.9: The handbook should include introductory material clarifying the possible roles of mechanistic and TK information in the IRIS assessment process. As it is recognized that methods and approaches in this area are continuing to evolve and be refined, the handbook need not (and cannot) specify every possibility but rather should focus on the most common roles for such data in the assessment process (see Table 2-1 of this report). EPA should obtain a professional editor's assistance (see Recommendation 2.1) in determining how best to organize the more detailed discussions of mechanistic TK information throughout the relevant sections of the handbook. *[Tier 2]*

Finding: The handbook recognizes the importance of assessing potentially susceptible populations by encouraging their consideration throughout the process. There are multiple places in the handbook that describe how the evidence base for susceptible populations should be incorporated into IRIS assessments. However, the handbook treats this area of hazard identification as a special case evaluation that may not be required unless animal, human, or mechanistic evidence points to a particular susceptibility or life stage. Discussions that define what constitutes evidence of susceptibility and describe the types of data that may inform such susceptibility are not provided in one place in the handbook, but those topics are alluded to in various places throughout the document.

Recommendation 2.10: The handbook should include introductory material summarizing how susceptible populations are to be identified, how relevant literature are to be sought and catalogued, and how this information is to be used to refine PECO statements and the assessment protocol. *[Tier 2]*

Finding: The handbook chapter on selection of studies for toxicity value determination does not directly follow on from the earlier handbook chapters. For example, it is unclear how the results of the systematic reviews that are conducted for hazard identification are used to select studies for the dose-response assessment. As described, hazard identification and toxicity value

determination appear to be disconnected processes. Thus, earlier portions of the handbook are difficult to evaluate without an adequately described connection to how studies will ultimately be selected for toxicity determination.

> **Recommendation 2.11:** The handbook should include introductory material describing the criteria for using the results of hazard identification and other considerations for the selection of studies for dose-response. *[Tier 2]*

Finding: The handbook lacks a section on the overall quality assurance of the IRIS assessment process.

> **Recommendation 2.12:** The handbook should include a section on ensuring the overall quality of the IRIS assessment process that establishes a minimum set of standards for quality assurance and provides procedures for monitoring quality and training EPA staff and contractors, as needed, to meet the quality assurance standards. *[Tier 2]*

Finding: Given the current complexity of the handbook and EPA's existing review process, it is difficult to see how the handbook could be updated in a timely manner. The handbook does not describe a process for its updating, which would include proposed timelines, how to identify major changes that would need to go through external peer review, and how to more quickly update details of evolving methods (e.g., study evaluation and software tools) that could be linked to the handbook without need for external review.

> **Recommendation 2.13:** The handbook should include content describing the timeline and process for updating of the handbook. *[Tier 2]*

Findings and Tier 3 Recommendation

Finding: Human epidemiology and animal toxicology studies are the main focus of the handbook. The handbook does not indicate the types of studies that might be considered for future IRIS assessments (e.g., high throughput, non-rodent vertebrate studies, and in silico). Consideration of chemical mixtures is a gap in the handbook that may be beyond the current scope of IRIS assessments. The committee urges EPA to plan for including additional types of studies in IRIS assessments.

Finding: Systematic reviews in environmental health are becoming more common, but the handbook does not describe how EPA will identify, evaluate, and incorporate systematic reviews from other sources.

> **Recommendation 2.14:** The handbook should include discussions that recognize new types of research and describe how EPA is preparing for possible inclusion of additional study types into IRIS assessments. To enhance the efficiency of the IRIS assessment process, EPA should consider how it will identify, evaluate, and incorporate systematic reviews from other sources, as they become more available. *[Tier 3]*

3

Planning Assessments

In this chapter, the committee reviews the portions of the chapters of the *ORD Staff Handbook for Developing IRIS Assessments* (the handbook) (EPA, 2020a) that describe processes related to planning the development of Integrated Risk Information System (IRIS) assessment:

- Chapter 1—Scoping of IRIS Assessments
- Chapter 2—Problem Formulation and Development of an Assessment Plan
- Chapter 3—Protocol Development for IRIS Systematic Reviews
- Chapter 4—Literature Search, Screening, and Inventory
- Chapter 5—Refined Evaluation Plan
- Chapter 7—Organizing the Hazard Review: Approach to Synthesis of Evidence
- Chapter 10—Analysis and Synthesis of Mechanistic Information

In reviewing those chapters, the committee considered whether the systematic review approaches used by the IRIS Program (outlined in Chapters 1-5 of the handbook) clearly describe and are consistent with methodologies considered to be state of the science (Question 2 in Appendix B), and whether the handbook clearly lays out a state-of-the-science approach for refinement of the scope and analyses of an IRIS assessment (Question 4 in Appendix B). The committee also assessed the handbook's process for evaluating and integrating mechanistic data and considered specific areas for improvement or recommended alternatives (Question 5 in Appendix B).

As requested by the U.S. Environmental Protection Agency (EPA), the committee also used planning documents for specific assessments (i.e., for Vanadium [EPA, 2021a] and Inorganic Mercury Salts [EPA, 2021b], the release of which spanned the release of the handbook).

OVERVIEW OF THE HANDBOOK'S MATERIAL ON PLANNING IRIS ASSESSMENTS

Chapter 1 "Scoping of IRIS Assessments" outlines the process for identifying the programmatic needs of EPA in relation to a proposed IRIS assessment. It presents examples of agencies and stakeholders with whom IRIS may need to engage in the process of defining the objectives of an assessment. The chapter also presents a series of five scope-related and eight prioritization-related questions to facilitate focusing the objectives of the assessment into the form of one or more population, exposure, comparator, outcome (PECO) statements. The scoping process overlaps with the problem formulation process in the next chapter of the handbook.

Chapter 2 "Problem Formulation and Development of an Assessment Plan" describes two processes: (1) the creation of a preliminary literature survey and (2) the drafting of an assessment plan. The assessment plan is a document that describes in broad terms how the assessment will

be conducted to address the programmatic needs determined by the scoping process. The preliminary literature survey identifies health effects that have been studied, as well as other types of information that the assessment may need to address. Confusingly, this chapter references Chapter 4 for conducting literature surveys, but Chapter 4 actually addresses a more systematic output referred to as a "literature inventory" (also referred to as a "systematic evidence map"). The literature inventory comprises data about populations, exposures, outcomes, study methods, and other study design features that are relevant to planning the specific systematic review questions and informing the other analyses that will deliver on the programmatic needs that are to be met by an assessment.

Chapter 3 "Protocol Development for IRIS Systematic Reviews" briefly describes how the IRIS program interprets the concept of a protocol for the conduct of a systematic review. Breaking with convention in systematic review terminology, the term "protocol," as used here, is a strategic document outlining an overall process for developing a planned approach to an overall IRIS assessment, rather than a detailed methodology for conducting a particular systematic review. In a presentation to the committee, EPA clarified that there are two "assessment protocols": an initial protocol that is published early and can change during the course of the assessment, and a final protocol that is published at the same time as the assessment and provides the full details on final methods that were chosen for conducting the systematic reviews and other elements of the assessment (Thayer, 2021). The initial protocol seems to be the subject of handbook Chapter 3; however, a template for the organization of what appears to be a final protocol is also presented.

Chapter 4 "Literature Search, Screening, and Inventory" provides guidance on conducting literature searches for IRIS assessments (Section 4.1), screening the results of search strategies for literature of actual relevance to an assessment (Section 4.2), and creating an inventory of the literature that has been determined to be relevant to an assessment (Section 4.3). Section 4.1 describes the use of bibliographic databases and methods for the development of search strategies for querying them. Section 4.2 describes the process of classifying ("tagging") studies as relevant or not to the assessment objectives, and for those that are relevant, adding further tags related to study design characteristics, topic area, and other aspects. The tagging process results in a pool of studies that meet the PECO criteria and a pool of "potentially relevant supplemental material," such as mechanistic studies, toxicokinetic (TK) studies, and pharmacokinetic or physiologically based pharmacokinetic modeling studies. Section 4.3 describes the process for creating the literature inventories, which is a database of the broad array of evidence relevant to the problem space as defined by the programmatic needs that were identified via the scoping process. The guidance on search and tagging explains the mechanics of building the literature inventory.

Chapter 5 "Refined Evaluation Plan" describes the process for moving from the assessment plan (handbook Chapter 2) and initial protocol (handbook Chapter 3 and as clarified by EPA [Thayer, 2021]) to a refined evaluation plan. The process is presented in the form of a series of questions intended to facilitate analysis of the data in the literature inventory in the context of the assessment plan, as well as development of additional targeted searches and analyses that were not presented in the assessment plan. The refined evaluation plan can be refined further into the final protocol.

Chapter 7 "Organizing the Hazard Review: Approach to Synthesis of Evidence" describes the planning of the hazard review, a specific product of the IRIS assessment process that presents conclusions about hazards from exposure to chemical substances. This process also

provides information that guides the dose-response analysis. The plan for the hazard review uses information obtained from more detailed analysis of the literature inventory, in particular the study evaluation process (handbook Chapter 6). The goal is to identify, prioritize, and group the set of endpoints on which the risk assessment is to be based and for which detailed data extraction will be conducted. The remaining endpoints are only summarized as supplemental material.

Chapter 10 "Analysis and Synthesis of Mechanistic Information" describes multiple activities related to the use of mechanistic information in the assessment process, including literature searching and screening, scoping and objective of analyses of mechanistic data, prioritization and evaluation of mechanistic studies, synthesis of mechanistic evidence, and use of mechanistic information in evidence integration and dose-response analysis. The material in handbook Chapter 10 is relevant to multiple chapters. As indicated in Question 5 (Appendix B), Sections 2.2, 4.3.3, and 6.6 of the handbook also discuss evaluation of mechanistic data.

RESPONSIVENESS TO THE 2014 NATIONAL ACADEMIES REPORT

The committee finds that the handbook is responsive to many of the planning-related recommendations made in the 2014 National Academies report *Review of EPA's Integrated Risk Information System (IRIS) Process* (NRC, 2014). In particular, through its initial literature survey, the development of a broad PECO statement, and subsequent development of a literature inventory based on a broad search, EPA has established a "transparent process for initially identifying all putative adverse outcomes," as recommended by NRC (2014, p. 37). Additionally, for all literature searches, the IRIS assessment process includes input from information specialists trained in conducting systematic reviews.

EPA's approach for using "guided expert judgment to identify the specific adverse outcomes to be investigated, each of which would then be subjected to systematic review," is partially responsive to the NRC (2014, p. 37) recommendation. While the handbook describes several steps for focusing IRIS assessments on specific health outcomes, as described further below, the process is not fully transparent in a manner consistent with conventional systematic reviews. Additionally, while the process outlined in the handbook is responsive to the NRC (2014, p. 37) recommendation to "include protocols for all systematic reviews conducted for a specific IRIS assessment as appendixes to the assessment," the committee found additional aspects of protocol development and reporting that could be improved.

CRITIQUE OF METHODS FOR PLANNING

A challenge in evaluating the processes described in the handbook is their presentation as being linear or funneled when many steps are in fact iterative. This is evidenced by the numerous call-forwards and call-backs between chapters (e.g., how Chapter 4 needs to be read to understand how to deliver the outputs of the processes described in Chapter 2, and the challenge in understanding where Chapter 3 fits in the planning process).

Therefore, in order to better communicate and render constructive its critique of the methods for planning IRIS assessments, the committee has divided the overall IRIS planning process into three stages, with a section of this chapter dedicated to each: problem formulation, protocol development, and organization of the hazard review (see Figure 3-1). Each stage has one or more accompanying processes, and each process is associated with one or more documented outputs relating to the overall assessment. Relatedly, addressing Question 4

(Appendix B), which asked whether the handbook lays out a state-of-the-science approach for refinement of the scope and analyses of an IRIS assessment, involved consideration of multiple processes described in handbook chapters 1, 2, 3, 4, 5 and 7. The various strengths and weaknesses of the processes regarding the state-of the-science are discussed throughout this chapter. However, it was not practicable to provide a summary response to the question.

This chapter also includes a discussion of the role of mechanistic and TK data in the planning process, and it covers special considerations for potentially susceptible populations in the planning of IRIS assessments. Although study evaluation is discussed in Chapter 4 of this report, it is included in Figure 3-1 to show the order of its presentation in the handbook.

Problem Formulation

The problem formulation process described in the handbook presents an important evolution in understanding how to use "literature inventories" or "systematic evidence maps" (databases of key features of scientific studies) to inform priority-setting in toxicity assessments. Literature inventories contribute to characterizing a general problem space by mapping existing relevant evidence (James et al., 2016). In the case of IRIS assessments, the space is defined by interpreting the problem into one or more broad PECO statements. The outcomes may be particularly broad, typically encompassing "all cancer and noncancer health outcomes." This contrasts with the types of PECO statements used in systematic reviews (including those developed by the National Institute of Environmental Health Science's Office of Health Assessment and Translation and using the Navigation Guide [Woodruff and Sutton, 2014] for environmental health), which are much narrower in scope.

The literature inventories are built using sensitive search strategies, the results of which are screened to remove off-topic research. The remaining studies are classified as to their likely role in the assessment. Features of the studies that are relevant to informing the objectives and approach of the assessment are extracted into a database. The results of this inventorying process help identify gaps, clusters, and patterns in a topic-relevant body of evidence, thereby informing decisions about when and how to conduct further primary or secondary research so as to fill critical gaps or synthesize important clusters. The creation of literature inventories is a key step for planning the assessment, and the general process described in the handbook reflects current best practice. This is to be expected, given the IRIS program's prominent role in developing these methods.

The creation and use of literature inventories constitute important recent innovations in making priority-setting a more evidence-informed process and improving the efficiency of chemical assessments. Identification of novel evidence clusters can be particularly important to inform a new or updated hazard conclusion or toxicity value. Efficiencies can be gained via avoiding the unnecessary analysis of evidence that is unlikely to result in such changes, instead focusing resources on decision-critical clusters. As such, the literature inventory is a key milestone in the IRIS process, incorporating information obtained from the scoping and initial problem formulation steps, and forming the foundation of all subsequent analysis. Although not cited in the handbook, these issues have been discussed in the academic literature prepared by scientists at the IRIS program and the National Toxicology Program (Walker et al., 2018; Wolffe et al., 2019; Keshava et al., 2020; Radke et al., 2020). The committee was somewhat surprised that the handbook did not make more reference to existing EPA research and other materials related to the planning process for IRIS assessments.

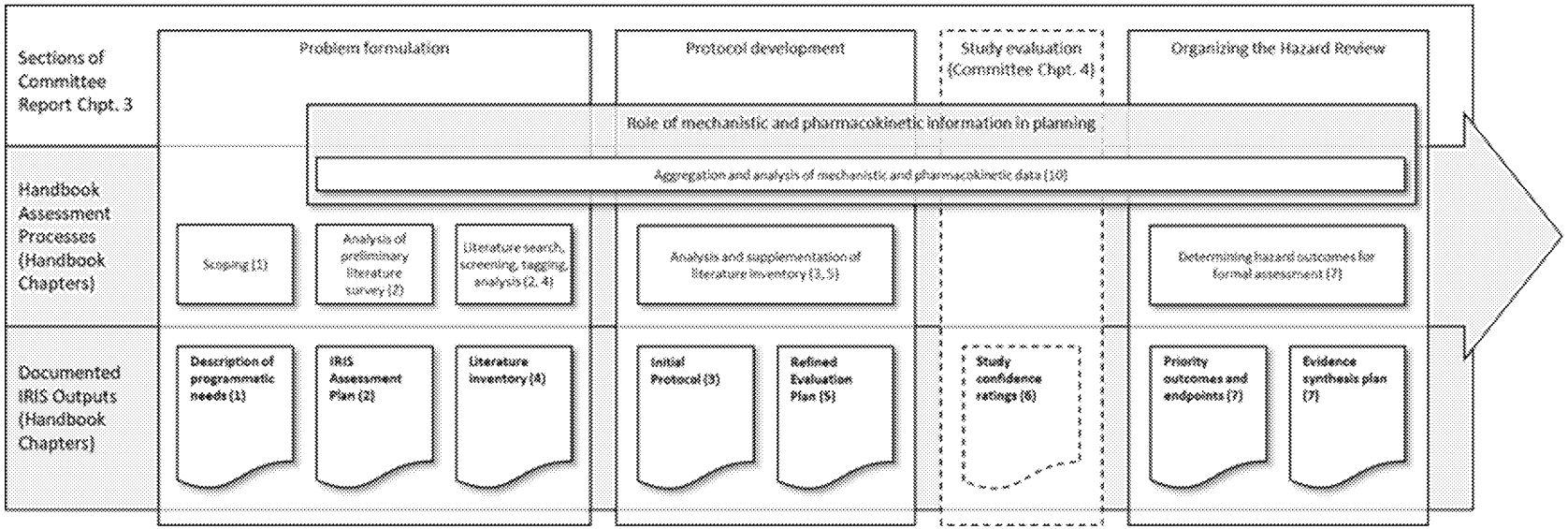


FIGURE 3-1 Mapping of three stages of the IRIS planning process identified by the committee.
Note: The figure illustrates the relationship of the stages discussed in Chapter 3 of this report with the processes and documented outputs described in the handbook. Study evaluation is shown in a dashed box because it is not covered in this chapter but is positioned before "Organizing the Hazard Review: Approach to Synthesis of Evidence" in the handbook.

The committee notes that literature inventories also have considerable potential value in and of themselves as a public good, providing a publicly accessible, queryable database that could be used by any research organization to identify knowledge gaps and clusters in toxicity research. Thus, literature inventories stand to increase not only the efficiency of EPA's work but also the efficiency of toxicity research, in general.

However, major challenges in understanding the processes described in the handbook are the inconsistent way in which they are documented, the lack of clarity about the order in which the steps are taken, and how the handbook chapters map onto these steps. For example, Figure O-2 (handbook p. xix) does not mention the assessment plan. It also positions the literature search and inventory development as occurring after the development of the systematic review protocol, when in fact these two elements have a fundamental role in the development of the protocol. As discussed later in this chapter, the protocol itself seems ambiguously characterized as a product of the planning process of an assessment, and in a manner at odds with conventional understanding of its use in systematic reviews.

One area that did not receive much attention in NRC (2014), but which has been a recurring issue in other National Academies reports, is the need to address susceptible populations and life stages. The handbook recognizes their importance by indicating that they should be considered in scoping and problem formulation (handbook Chapters 1 and 2); PECO statements consider the need for a specific PECO question for susceptible populations (handbook Chapter 3); the assessment plan (handbook Chapter 3) and refined assessment plan (handbook Chapter 5) take into account susceptible populations; mechanistic information is to be used to assess the types of susceptibility that may exist (handbook Chapter 10); and evidence synthesis, integration, and dose-response all consider the possibility of susceptible populations (handbook Chapters 9, 11, 12, and 13). The committee notes that the literature inventory appears to be an ideal step during which a systematic search for information on susceptible populations and life stages could be performed. However, there is little mention of this issue in either the search strategy development or the discussion of literature tagging. Although susceptibility is mentioned in the discussion of supplemental literature searches in Chapter 4 of the handbook, there is an opportunity to more explicitly address this issue during the implementation of the broad literature search underlying the literature inventory, for instance through the use of tags reflecting categories of susceptibility identified in handbook Table 9-2 (p. 9-6).

Finally, a critical aspect of problem formulation is the identification of "key science issues" that the assessment will need to address. Some of the issues that the handbook mentions include conflicting data, adversity of endpoint, human relevance, and differential susceptibilities. However, the handbook does not provide a systematic process for identifying such issues. It appears that they are largely identified through the preliminary literature survey, stakeholder engagement, and public comment. Systematic stakeholder engagement and issue identification are themselves areas of active research for which no broadly accepted methodology currently exists. Some of the key concerns with existing stakeholder engagement and issue identification include the disparity in resources available to different stakeholders and bias in issue identification due to vested interests (Haddaway and Crowe, 2018).

The committee's findings and recommendations regarding problem formulation are presented in the final section of this chapter.

Protocol Development

This section discusses the processes described in the handbook for developing and making publicly available, in the form of a protocol, the planned objectives and methods for conducting an assessment.

In a systematic review, the protocol is a complete account of planned methods that should be registered (i.e., publicly released in a time-stamped, read-only format). A protocol can be registered in a specialist protocol repository, on a preprint server, or via formal publication in a scientific journal. In some circumstances, organizations (e.g., public agencies) may register their protocols via publication on a website in a manner that safeguards against changes without the reader being aware of it. The committee considers EPA to be an organization that could employ publication via its public website as a means of protocol registration. Some advantages of protocol registration include discouraging the shaping of methods around expected findings, as methods are defined and published prior to data extraction and analysis; and rendering transparent any changes in planned methods that might affect the integrity of the review process and credibility of its findings (Nosek and Lakens, 2014).

In the handbook, the term "protocol" refers to as many as three document types: the "initial" or "draft" protocol that is the product of the problem formulation process (Chapter 3); the refined evaluation plan that is to be the product of a protocol formulation process yet is also described as being "in the protocol" (Chapter 5); and a "final protocol," which has no dedicated chapter but is published alongside the final assessment. Each document is different but their exact content, and how a later document builds on a previous one, is unclear.

The scope of each of these documents is also ambiguously described in the handbook. For example, the document described as a protocol in Chapter 3 seems to be intended as a broad plan for conducting a systematic review that delivers the objectives defined in the assessment plan, specifying the approaches and analyses for the entire IRIS assessment process, not just the systematic review portion. Additionally, when it is initially developed (prior to Literature Search), the steps subsequent to the literature inventory are at this stage fairly generic. While this protocol includes a broad PECO statement, the summary box "Organization of the Protocol" on page 3-3 of the handbook presents a level of detail more readily associated with a formal systematic review protocol and seems to better match the intent of the final protocol. To further complicate matters, EPA presented the IRIS protocol as consisting of two documents, an initial and final protocol, both of which seem potentially open to revision in the course of conducting an IRIS assessment (Thayer, 2021, slide 8).

The committee interpreted the handbook to indicate that there is one protocol that is revised over time, accumulating focus and detail as the scope of the problem and the plan for how to solve it using systematic reviews and other analyses evolves throughout the assessment process. This increase in focus seems reasonable given the potential scope and complexity of IRIS assessments in the context of the programmatic needs of EPA, and evolution of protocols is usually expected. However, the different names given to these documents and the unclear differentiation of each stage of the process documented in the handbook makes it very challenging to discern what methods are being documented, when, and to what purpose.

Of the three protocols or planning documents described in the handbook, the final protocol is the most consistent with a conventional definition of a systematic review protocol as being a complete record of planned methods. It is notable, however, that this plan is only published alongside the final assessment rather than in advance of it. Consequently, IRIS assessments can accrue only some of the benefits of protocol publication, such as improving

planned methods in response to external feedback. Critically, the IRIS approach to protocol publication cannot be used to audit changes to planned methods, as these methods are not being publicly documented in a way that makes it possible to determine if any changes to the protocol are made before or after conducting the analysis. Other government agencies have registered protocols (e.g., EFSA et al., 2017) and, as noted above, publication of the protocol via the EPA public website could potentially be considered equivalent to such third-party registration.

Additionally, the processes for inclusion and exclusion of literature from the systematic reviews conducted as part of IRIS assessments diverges from current best practices. In conventional systematic reviews, the inclusion and exclusion of evidence, as well as delineation of units of analysis for evidence synthesis and strength of evidence conclusions, are strictly governed by prespecified PECO statements, in order to be fully transparent and minimize potential for bias via selective inclusion of literature in a review. For IRIS assessments, a unit of analysis could be defined at the endpoint level (e.g., clinical chemistry) or outcome level (e.g., liver toxicity). However, the handbook as well as recently released planning documents (e.g., for Vanadium [EPA, 2021a] and Inorganic Mercury Salts [EPA, 2021b]) all only include broad PECO statements, along with broad health effect categories. This approach contrasts with conventional systematic reviews, in which even within a relatively broad health effect category (e.g., cardiovascular disease), a detailed PECO statement particular to that health effect is prespecified so as to ensure that issues relating to differing target populations, windows of susceptibility, and measures of outcome can be addressed up front. Moreover, as currently described in the handbook, evidence may be triaged (excluded from further consideration) at the "refined plan" step (handbook Chapter 5) or after study evaluation at the "organize hazard review" step (handbook Chapter 7). The reasons for the triage of such evidence, and safeguards to ensure that evidence is not being used selectively, are not sufficiently clear in the handbook. The development of refined PECO statements for each health effect category at the assessment protocol stage would therefore provide clarity as to how specific pieces of evidence are included or excluded, as well as transparency as to any subsequent changes to those criteria.

The committee recognizes that documenting the planning process for IRIS assessments is complex. It is clear that the IRIS process is not linear, but iterative, and that at several points the planned methods are revisited, revised, and refined. It may be possible to simplify the process and number of associated documents and state more clearly where and when planned processes are revised. One possible approach is to have two planning documents: (1) an assessment plan that is a broad statement of the overall problem and approach being taken to the assessment, and (2) an assessment protocol that describes in detail the specific methods to be used to conduct hazard identification and dose-response assessment. The assessment protocol would be developed based on information available from the literature inventory and could be subsequently revisited and refined (e.g., after public comment), with revisions or refinements documented and the final draft of the assessment protocol registered in advance of the detailed analysis of the hazard identification and dose-response stages of the assessment. Any deviations from the planned methods described in the registered assessment protocol can be documented when the draft IRIS assessment is released for public comment and peer review. For instance, the handbook (in the "Organizing the Hazard Review: Approach to Synthesis of Evidence" chapter) currently appears to suggest that changes as to what outcomes and endpoints are synthesized may be made after study evaluation. These are the changes from the registered protocol that would need to be documented.

Figure 3-2 and Table 3-1 illustrate how the planning documents and literature inventory could be related, consistent with the committee's understanding of the Vanadium planning documents. Specifically, a broad PECO statement that was used to develop a literature inventory was included in the Vanadium planning documents (EPA, 2020b) released on July 1, 2020, for public comment, and the results of this inventory were subsequently described in the Systematic Review Protocol (EPA, 2021a) released for public comment in April 2021. However, two aspects of protocol development that are typically included in conventional systematic reviews are missing from the process for Vanadium: (1) registration of the protocol (after addressing public comments and prior to conducting analysis) and (2) inclusion of refined PECO statements to guide the unit of analysis at which evidence synthesis and strength of evidence judgments are conducted.

The committee's findings and recommendations regarding protocol development are presented in the final section of this chapter.

TABLE 3-1 Illustrative Documentation of an Assessment Plan, Literature Inventory, and Assessment Protocol from the IRIS Planning Process[a]

Planning Product	Contents	Initial Version	Subsequent Versions
Assessment Plan	• Scoping/problem formulation • Preliminary literature survey • Science issues • Methods for assembling literature inventory (including broad PECO, search/screening strategies)	Draft for public comment, published after scoping and problem formulation stages	• Revised version incorporating changes based on public comments and results of literature inventory published with draft assessment protocol
Literature Inventory	Study database and categorization • Health effects • Susceptible populations • Mechanistic and TK information	Initial inventory made publicly available alongside assessment protocol draft for public comment	• Updated periodically (not for public comment) • Inventory for public comment available with public comment/peer review draft of IRIS assessment • Final inventory published alongside final IRIS assessment
Assessment Protocol	• Systematic review protocols (including refined PECOs specifying units of analysis; methods for study evaluation and analysis) • Evidence synthesis and integration methods (including use of mechanistic and TK information) • Dose-response methods	Draft for public comment, published alongside initial literature inventory	• Revised version, incorporating changes based on public comments, registered prior to commencement of detailed analyses • Version detailing the actual approach taken with explanation of deviations from registered methods, published alongside public comment and peer review draft of IRIS assessments • Final version, incorporating any changes based on public comment and peer review, published alongside final IRIS assessment

[a] The table indicates the content of each suggested planning product illustrated in Figure 3-2, when each product might first be made available, the purpose of each product, and when each product would be revised. This is an illustration of a possible approach and is not a recommendation as to how the IRIS assessment should be documented.

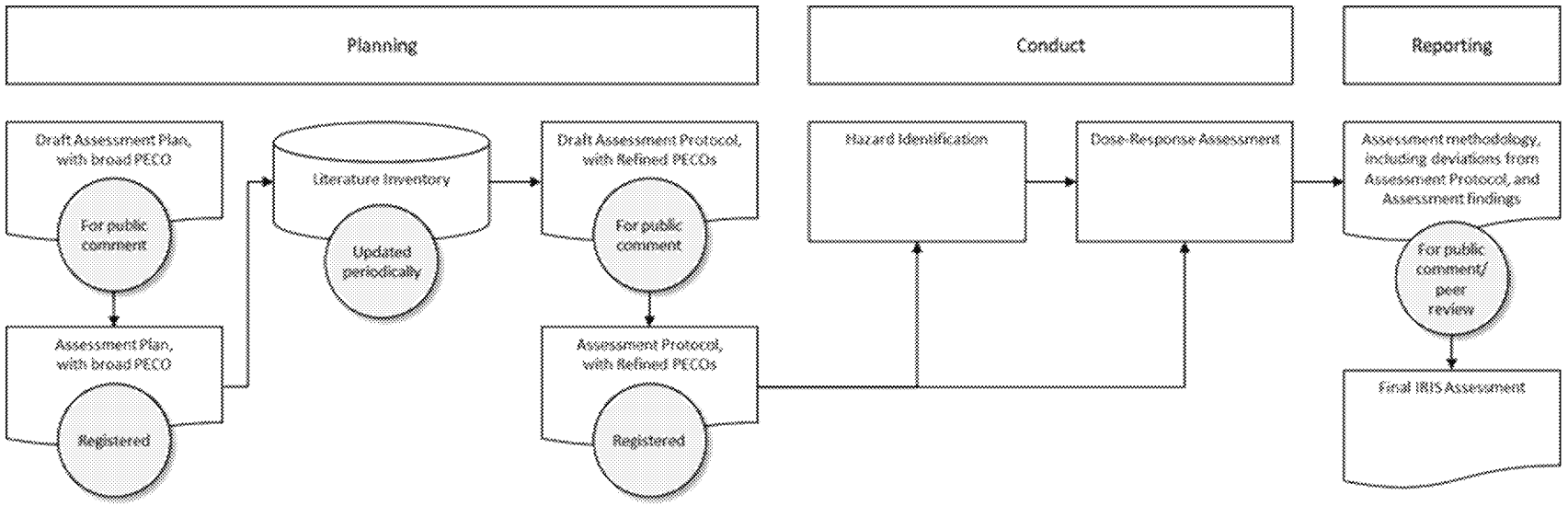


FIGURE 3-2 Illustration of an assessment plan, literature inventory, and assessment protocol from the IRIS planning process.
Note: The figure illustrates which of the three products might be updated or registered, and where the products would feed into the IRIS assessment process. It is important to note that this is an illustration of a possible approach and is not a committee recommendation as to how the IRIS assessment should be documented.

Organization of the Hazard Review

The intent and scope of "Organizing the Hazard Review: Approach to Synthesis of Evidence" (handbook Chapter 7) were not clearly presented in the handbook. However, given that this step is presented after study evaluation, it appears that a main goal is to use judgments of study confidence to further identify and prioritize the most relevant endpoints for hazard identification. Of concern, however, is that criteria beyond study confidence are included in the text of Chapter 7 and Table 7-1 (p. 7-4). For example, study outcomes including evidence of greater sensitivity and dose-response are listed as elements to consider during the organizational process. Also page 7-3 of the handbook states, "These questions extend from considerations and decisions made during development of the refined evaluation plan to include review of the uncertainties raised during individual study evaluations as well as the direction and magnitude of the study-specific results" (EPA, 2020a). Thus, this step appears to be incorporating criteria other than study confidence to organize the hazard review. As such, this step would serve as another opportunity for refining the protocol prior to data extraction and synthesis. If that is the case, this purpose would need to be explicitly stated and the process thoroughly documented. Significantly, it should be noted that inclusion of this step deviates from typical systematic review approaches.

As written, the organizational process is poorly defined, too open-ended, and too vulnerable to subjectivity. Importantly, as other criteria should have been addressed in earlier steps, a compelling justification would be needed if any criteria other than study confidence are used to exclude data at this late stage of the assessment process. Ideally, the organization and planning of the hazard review are very narrow in scope and primarily focused on evaluating how well the identified data pool can address the PECO questions. If, instead, this step is effectively an opportunity for further refinement of the protocol or the PECO questions themselves prior to data extraction, that would need to be made explicitly clear and those refinements would need to be published and time stamped in the same way as was done for similar steps in the planning process.

Critically, the organization and refinement processes lack specific criteria to ensure transparency and reproducibility, and are not constructed to minimize subjectivity to the strongest degree possible. Although it makes sense that the outcomes and hazards with the strongest evidence should be considered first, Chapter 7 of the handbook does not operationalize how the strength of evidence would be established or ranked. Instead, it suggests questions to be potentially considered throughout the decision process. For example, the use of "probably" in the phrase "if the studies are all *low* confidence due to reduced sensitivity, the outcome should probably be summarized" (EPA, 2020a, p. 7-2) leaves the process too vague and vulnerable to subjectivity. Many of the questions should have been sufficiently addressed far earlier in the process including at what level the outcomes will be grouped. Life stage, exposure duration, and exposure route should have been factored into the crafting of the PECO statements. Dose-response would be better handled after the data extraction. It is inappropriate to use study outcomes to inform data extraction, because doing so would be a significant deviation from systematic review procedures.

Throughout Chapter 7 of the handbook, as written, the individual processes are confusing and do not provide a clear path to operationalization. For example, in the first paragraph when acknowledging that human and animal evidence do not always align the text states that "decisions can sometimes be informed by specific mechanistic evaluations, for example analysis of the extent of the linkage between related outcomes" (EPA, 2020a, p. 7-1). It is unclear what is meant by that statement, what mechanistic data or evaluations are included, or what is meant by

"extent of the linkage." If the goal is to identify endpoints in animals that are informative for a specific endpoint in humans (the example given in the text is spontaneous abortion), those linked endpoints would need to be stated earlier in the process (if those relationships are known) or reserved for the evidence integration step when the collective set of data are being interpreted.

It is also unclear when and how additional mechanistic data would enter the data organization process. For example, page 7-3 of the handbook states that "additional targeted searches for mechanistic information specific to those health effects and/or organ systems may need to be performed. These supplemental searches may involve new literature search strategies, and they may be health effect- or tissue-specific rather than chemical-specific" (EPA, 2020a). Explicit criteria for when those searches may be needed, how those searches should be conducted, and how the data will be incorporated into subsequent steps need to be provided.

The committee's findings and recommendations regarding organizing the hazard review are presented in the final section of this chapter.

Role of Mechanistic and Toxicokinetic Information in Planning

As discussed in Chapter 2 of this report, there are multiple places where the roles of mechanistic and TK data are described in the handbook, and they are not entirely consistent. Moreover, the multiple instances of both backward and forward referencing make the handbook confusing as to when and how mechanistic and TK data are incorporated into the assessment. As described in NASEM (2017a), mechanistic and TK data can be very informative during the scoping and problem formulation phase, helping determine what outcomes to focus on and how animal and human evidence could be integrated. For instance, in its systematic review of phthalates and male reproductive tract development, NASEM (2017a) focused on endpoints relevant to the anti-androgenic activity of phthalates, such as fetal testosterone, anogenital distance, and hypospadias, and did not include cryptorchidism in this grouping of endpoints. Thus, development of refined PECO statements can be informed by mechanistic and TK considerations with respect to which endpoints are grouped together, and which are grouped separately. It is recognized, however, that during this planning stage, a systematic and critical evaluation of mechanistic and TK data has not yet occurred, and that the IRIS program will need to rely on expert judgment and input from public comments in utilizing such information in the planning process.

In other cases, mechanistic data may warrant their own PECO statement (or set of PECO statements) because it is appropriate and possible to evaluate their strength of evidence using systematic review methods. In the example provided in EPA's presentation, the mutagenicity of chloroform was a case in which a separate PECO statement was developed to inform analysis of a mutagenic mode of action (MOA) for cancer (Thayer, 2021). It is also possible that certain mechanistic endpoints that are considered to be precursors to apical endpoints or adverse effects warrant specification as a discrete unit of analysis, particularly if such endpoints are eligible for use in dose-response assessment. The handbook does not provide much guidance during the planning stages as to how these decisions are made.

With respect to the role of mechanistic data in planning IRIS assessments, the use of key characteristics (KCs) for searching and organizing mechanistic data for informing hazard identification is unclear in the handbook. However, it was clear from EPA's presentation that the intention is to use KCs as they are available for different health effects to help organize

mechanistic information (Thayer, 2021). KCs comprise the set of chemical and biological properties of agents that cause a particular toxic outcome, and were introduced first for carcinogens (Smith et al., 2016) and subsequently for several other types of toxicants, including male and female reproductive toxicants, endocrine disrupting chemicals, hepatotoxicants, and cardiotoxicants (Arzuaga et al., 2019; Luderer et al., 2019; La Merrill et al., 2020; Lind et al., 2021; Rusyn et al., 2021). It is generally agreed that KCs can be used for screening and organizing mechanistic data, in line with their originally stated purpose in Smith et al. (2016). As a result, KCs have been incorporated into hazard assessment approaches by several organizations (NTP, 2018a,b; IARC, 2019; Samet et al., 2020). On the other hand, there have been concerns that use of KCs will lead to an increase in false positives (Smith et al., 2021) and that they may not be predictive of cancer hazard, especially as some KCs such as oxidative stress lack specificity for cancer and other outcomes (Becker et al., 2017; Goodman and Lynch, 2017; Guyton et al., 2018; Smith et al., 2021). However, the approach used for identifying KCs explicitly emphasizes "sensitivity" rather than "specificity." In particular, KCs are identified through a process that examines known toxicants for a particular outcome (e.g., cancer) and represents common characteristics of those toxicants. Thus, the logical conclusion is that "if X causes outcome Y, then X has one or more KCs." On the other hand, a predictive statement would be the converse: "If X has one or more KCs, then X causes outcome Y." However, from elementary logic, a true statement does not directly imply that its converse is also true—only that its contrapositive is true. The corresponding contrapositive is "if X does NOT have one or more KCs, then X does NOT cause outcome Y." Further research or other approaches are needed to make predictive statements about toxic outcomes.

This line of reasoning suggests that from a hazard identification perspective, KCs may be useful in the following senses if a toxicant X has been adequately investigated for all KCs:

- If toxicant X exhibits one or more KCs, it is biologically plausible that X causes outcome Y.
- If toxicant X exhibits none of the KCs, then it is biologically implausible that toxicant X causes outcome Y (unless a new KC has been identified).

Here, the committee understands the term "biological plausibility" in the sense of being consistent with existing biological knowledge, but not necessarily having an established causal mechanism. Put another way, toxicant X exhibiting one or more KCs is necessary, but not sufficient, for toxicant X to cause outcome Y. Used in this manner, evaluating whether an agent exhibits KCs and identifying data gaps as to what KCs have or have not been sufficiently studied can be informative for hazard identification, even independent of further analysis in the form of MOAs or adverse outcome pathways. Thus, a key benefit of organizing mechanistic data using KCs is that they can be readily used in evidence integration to transparently address biological plausibility. As discussed in NASEM (2017b), p. 121), the main advantage of this approach is that it "avoids a narrow focus on specific pathways and hypotheses and provides for a broad, holistic consideration of the mechanistic evidence." As a result, their use promotes consistent approaches across chemicals and avoids only "looking under the lamppost." These advantages are consistent with the purpose of the handbook, and thus it is appropriate for EPA to use KCs in this manner.

FINDINGS AND RECOMMENDATIONS

Findings and Recommendations Related to Problem Formulation

Finding and Tier 1 Recommendation

Finding: The process of developing the literature inventory is in accordance with established best practices in searching for and screening evidence, starting with a comprehensive search to build a database that can then be interrogated for refining the assessment objectives. It makes suitable use of information specialists and is notable for its comprehensiveness of coverage of both published and grey literature resources.

> **Recommendation 3.1:** The handbook should make explicit which components of a literature inventory database are to be made publicly available and when. *[Tier 1]*

Findings and Tier 2 Recommendations

Finding: The relationship among the scoping process, the development of the literature inventory, and the analysis of the inventory to define the assessment plan is broadly appropriate and consistent with an understanding of best practice. Overall, EPA's process is responsive to the recommendation from NRC (2014) that assessments identify all putative adverse outcomes and then prioritize them for systematic review. However, little of the relevant literature in this area, including publications by EPA scientists, is cited to provide support for the methods outlined in the handbook.

> **Recommendation 3.2:** The handbook should more fully reference the literature in environmental health and other disciplines that relates to evidence mapping and its role in planning research, in order to strengthen the rationale for the method and provide relevant citations. *[Tier 2]*

Finding: The planning sections of the handbook, in particular the development of the literature inventory, could more directly support the identification of information and gaps with respect to susceptible populations and life stages. Adding a component to the literature inventory addressing this issue can be one way to better compile, organize, and assess the information available. This could help address the crosscutting question that each chemical assessment faces of whether the evidence base is adequate to address the possibility of susceptible populations and life stages and whether that evidence points to unique susceptibilities and/or particular uncertainties. Having this information identified in the literature inventory could support the refinement of the assessment plan and later steps in the process, including dose-response assessment.

> **Recommendation 3.3:** The literature inventory process should consider tagging studies for the presence of data pertinent to identification of susceptible populations. *[Tier 2]*

Finding and Tier 3 Recommendation

Finding: Identification of key science issues remains an area of problem formulation where systematic methods are needed, but broadly accepted approaches that address potential for bias have not yet been developed.

> **Recommendation 3.4:** EPA should consider developing a more systematic approach to identifying science issues. *[Tier 3]*

Findings and Recommendations Related to Protocol Development

Finding: Recognition in the handbook of the importance of publishing protocols is welcome. This responds to a key recommendation of the 2014 NRC Report.

Finding: Appropriate documentation of the planned methods requires refinement for consistency with best practice relating to protocol development and publication in systematic review. Consistent with the recommendation from NRC (2014), the handbook specifies publication of the final protocol as appendixes to the final assessment. However, the lack of any earlier public documentation of the protocol, and any modifications, is an important shortcoming because it does not sufficiently ensure the transparency and credibility of the methods and findings of the assessments.

Finding: The handbook lacks clarity regarding the products of the planning process, the relationships among them, which are expected to be updated or should be registered, and how they feed into the IRIS assessment.

> **Recommendation 3.5:** The handbook should clarify and simplify the assessment planning process as follows: restructure the handbook to directly reflect the order in which each step is undertaken, unambiguously identify each of the products of the planning process, clearly define what each product consists of, and state if and when each product is to be made publicly available. *[Tier 1]*

> **Recommendation 3.6:** EPA should create a time-stamped, read-only final version of each document that details the planned methods for an IRIS assessment prior to conducting the assessment. *[Tier 1]*

Finding: The handbook diverges from the best practices in systematic reviews in that the unit of analysis for evidence synthesis and strength of evidence conclusions is not clearly defined in refined PECO statement(s). The handbook includes a "broad PECO," intended to support the literature inventory. However, it does not specify the processes for developing more refined PECO statements for each unit of analysis, which could be defined at the endpoint level (e.g., clinical chemistry) or outcome level (e.g., liver toxicity). PECO statements in clinical systematic reviews often employ tiering of outcomes, for instance using hierarchies of broader to narrower categories, distinguishing between "critical" and "important" outcomes, and flexibility to group outcomes (e.g., see Higgins and Thomas, 2019, Section 3.2.4.3). Similar approaches can be employed by the IRIS program to organize endpoints and outcomes. It is recognized, however,

that during this planning stage, a systematic and critical evaluation of mechanistic and TK data has not yet occurred, and that the IRIS program will need to rely on expert judgment and input from public comments in utilizing such information in the planning process.

> **Recommendation 3.7:** The IRIS assessment protocol should include refined PECO statements for each unit of analysis defined at the level of endpoint or health outcome. The development of the refined PECO statements could benefit from considerations of available mechanistic data (e.g., grouping together causally linked endpoints, separating animal evidence by species or strain) and TK information (e.g., grouping or separating evidence by route of exposure), and information about population susceptibility. *[Tier 1]*

Findings and Recommendation Related to the Organization of the Hazard Review

Finding: Explicit consideration needs to be given to the organization of the hazard review in the planning stages of the IRIS assessment, but the handbook leaves this planning process under-specified. This is problematic, given that this step represents the highest level of granularity in describing how evidence is to be selected, organized, and grouped going into the data extraction stage of an assessment. For transparency, there needs to be enough detail that evidence synthesis and integration can be readily mapped back to the planned methods.

Finding: Although the "organize hazard review" step is positioned in the handbook after the study evaluation stage of an assessment, it in fact seems that most of the questions being posed in Chapter 7 of the handbook can be answered after the assembly of the literature inventory. The exception is the prioritization of study groups based on the results of the study evaluation step. More clarity about what is being done at this stage of the planning process, in particular which elements would more appropriately be subsumed under other (earlier) stages, is needed.

Finding: This step deviates from best practices in conventional systematic reviews, where all outcomes are prespecified and not subject to change after study evaluation. The handbook does not provide sufficient justification for revisiting the design of the systematic review after study evaluation. In particular, it does not articulate clearly what findings from study evaluation would be sufficient to change the analysis plan.

> **Recommendation 3.8:** The steps of organizing the hazard review should be narrowed to focus on new information obtained after the study evaluation stage. Organizing the hazard review should be structured by clear criteria for triage and prioritization, and aimed at producing transparent documentation of how and why outcomes and measures are being organized for synthesis. *[Tier 1]*

Findings and Recommendations Related to Mechanistic and Toxicokinetic Data

Finding: Mechanistic and TK data are potentially highly informative during the planning and protocol development process. The current handbook describes these roles in multiple places, and they are not entirely consistent.

Finding: The handbook is not clear as to when mechanistic and TK data may require a separate PECO statement defining a discrete unit of analysis for systematic review, synthesis, and strength of evidence judgments.

> **Recommendation 3.9:** The handbook should describe how the IRIS assessment plan and IRIS assessment protocol can identify the potential roles of mechanistic and TK data, including if they are to be units of analysis for systematic review, synthesis, and strength of evidence judgments. At a minimum, all endpoints that may be used for toxicity values, including so-called "precursor" endpoints that might be viewed as "mechanistic," should require separate PECO statements; however, application of systematic review methods to other mechanistic endpoints, such as mutagenicity, may depend on the needs of the assessment. The key mechanistic and TK questions should be identified to the extent possible in the IRIS assessment plan and IRIS assessment protocol. *[Tier 1]*

Finding: The use of KCs to search, screen, and organize mechanistic data is increasingly accepted and was highlighted in EPA's presentation to the committee.

Finding: The role of KCs for informing hazard identification has been the subject of debate. They are appropriate for use in evaluating biological plausibility, or lack thereof. However, KCs as currently constructed tend to be sensitive but not necessarily specific. More research is needed into whether and how they can be used to be more predictive of hazard.

> **Recommendation 3.10:** When available, KCs should be used to search for and organize mechanistic data, identify data gaps, and evaluate biological plausibility. Those uses should be reflected in the IRIS assessment plan and IRIS assessment protocol. *[Tier 1]*

4

Study Evaluation

The committee reviewed "Study Evaluation," Chapter 6 of the *ORD Staff Handbook for Developing IRIS Assessments* (the handbook) (EPA, 2020a). The committee's review of that chapter considered the adequacy of the study evaluation methods presented for individual human studies (epidemiology and controlled exposure), animal studies, mechanistic evidence (pilot testing approaches), and pharmacokinetic (PK) models. It also considered how those methods might be improved (Question 3 in Appendix B).

OVERVIEW OF THE HANDBOOK'S MATERIAL ON STUDY EVALUATION

Chapter 6 of the handbook describes approaches for evaluating individual studies for their validity and utility in assessing a health effect associated with specific exposure/endpoint relationships, with primary focus on study utility for hazard identification. The focus is on human epidemiological and experimental animal toxicology studies, with additional considerations for controlled human exposure studies, computational PK models and physiologically based pharmacokinetic (PBPK) models, and data relevant to mechanisms of toxicity. The stated goal of study evaluation is to evaluate "the extent to which the results are likely to represent a reliable, sensitive, and informative presentation of a true response" (EPA, 2020a, p.6-2). While standard methods for systematic review would consider only risk of bias when evaluating individual studies (Higgins and Thomas, 2019), study evaluation in the handbook is intentionally broadened to include the additional study evaluation domains of "sensitivity" and "reporting quality." It is not standard practice to include the concepts of sensitivity or reporting quality as part of the evaluation of individual studies included in systematic reviews of human research, although these domains are sometimes part of study evaluation for systematic reviews of animal studies (SYRCLE [Hooijmans et al., 2018). For the human studies, evaluation for risk of bias and sensitivity follows the Risk of Bias in Non-randomized Studies of Interventions (ROBINS-I) framework (Sterne et al., 2016), with a modification to include a domain for sensitivity and accommodate types of studies typically encountered in environmental and occupational epidemiology. Human studies also consider reporting quality as a contributing feature to the risk of bias and sensitivity domain judgments. Evaluation of animal toxicological studies assess risk of bias, sensitivity, and reporting quality as separate study evaluation concerns. Individual evaluation domains differ slightly to accommodate common differences across human, animal, and controlled exposure studies, but in all cases domain-specific judgments are combined toward an "overall rating" of study confidence that summarizes the evaluation for a specific exposure/endpoint relationship from a given study as "high," "medium," "low," "or "uninformative." Such rating guides the relative weight that each study will receive in downstream hazard identification, with "low confidence" studies generally not used as primary sources of information, and "uninformative" studies not considered further in evidence synthesis or integration. As mentioned previously, Chapter 6 of the handbook states on page 6-4 that study evaluations are "generally" conducted independently

by at least two reviewers, which leaves some indication of an undesirable possibility that only one reviewer would evaluate a study.

RESPONSIVENESS TO PREVIOUS NATIONAL ACADEMIES REPORTS

The committee finds that the handbook is responsive to many of the recommendations provided in the 2014 National Academies report *Review of EPA's Integrated Risk Information System (IRIS) Process* relative to evidence evaluation, including that the U.S. Environmental Protection Agency (EPA) continue to adapt or develop tools for assessing risk of bias in different types of studies, that risk of bias assessments should be incorporated into the IRIS process and carried forward to the evidence integration step, and that EPA should publish its risk of bias assessments. The 2014 report noted that EPA could augment risk of bias considerations with additional "quality assessment items" that are relevant to a particular systematic review if they are empirically based, as may be the case in the handbook with consideration of sensitivity and reporting quality.

The handbook's use of study evaluation ratings to exclude studies from further consideration is only partially concordant with the 2014 report, and recommended practice for using study evaluation to exclude studies prior to data extraction has evolved since the 2014 report. The 2014 report stated, "The risk-of-bias assessment can be used to exclude studies from a systematic review or can be incorporated qualitatively or quantitatively into the review results. The plan for incorporating the risk-of-bias assessment into a systematic review should be specified a priori in the review protocol" (NRC, 2014, p. 76). It was a clear expectation, however, that study exclusion be a rare and transparent process and only studies with severe methodological shortcomings ("fatal flaws") be excluded. Importantly, the 2014 report's recommendations pertained to the exclusion of studies based on risk of bias assessment and not on additional quality assessment items such as sensitivity and reporting quality. The handbook's reliance on these additional quality assessment items, and their possible use for study exclusion, represents a potential departure from the recommendations of the 2014 report. In addition, other National Academies reports that followed the 2014 report have recommended that EPA not use the risk of bias analysis to exclude studies. For example, the 2021 National Academies report *The Use of Systematic Review in EPA's Toxic Substances Control Act Risk Evaluations* explicitly recommends that EPA "not exclude studies based on risk of bias, study quality, or reporting quality" (NASEM, 2021, p. 40).

Because financial ties of the investigators can be strongly associated with favorable outcomes for the sponsors, even when taking into account other risks of bias, NRC (2014) also recommended that funding sources be considered in the risk-of-bias assessment for individual studies included in systematic reviews that are part of an IRIS assessment. That recommendation was based largely on evidence obtained from clinical studies because less was known at that time about the extent of funding bias in animal research. Since the 2014 report was published, evidence for funding bias in both the preclinical human and animal literature has increased, as discussed in Chapter 2 of this report. This recommendation has not been addressed.

CRITIQUE OF METHODS FOR STUDY EVALUATION

The committee's critique of study evaluation methods in the handbook focuses mainly on human epidemiological and experimental animal toxicology studies. Evaluations of PBPK and

PK models and data relevant to mechanisms of toxicity are discussed at the end of this critique section. The objectives of handbook Chapter 6 are clear and represent an important step in the IRIS assessment process. Furthermore, study evaluation is a critical step in systematic review methods. While much of the study evaluation process is clear and well defined, the committee identifies two major areas where a lack of clarity in the handbook is likely to inhibit operationalization, transparency, and reproducibility of the study evaluation process. The first area relates to the description of how studies evaluated to have certain confidence ratings will be used in future steps of the IRIS assessment. The second, more substantive, area relates to a lack of clear delineations among the separate concepts of risk of bias, reporting quality, study sensitivity, and other standard concepts in systematic review. The lack of clarity is particularly complicated in the context of IRIS assessments, due to the apparent differences in how study evaluation has historically occurred for human studies and animal and toxicological studies. While Thayer (2021) and EPA's written responses to the committee's questions provided important clarifications in both of these areas and left the committee confident that IRIS staff are evaluating studies appropriately, deficiencies and ambiguities in the text of the handbook threaten the ability to guide transparent, credible reviews of the evidence.

How Study Confidence Ratings Are Used Throughout the IRIS Assessment Process

The handbook explicitly states that study evaluation is performed in the context of a study's utility for identification of individual hazards, rather than the usability for dose-response analysis. The committee notes confusion about how studies' confidence ratings will be used toward this purpose. A major concern is that studies that are judged as "critically deficient" or "deficient" in one evaluation domain are typically rated as "uninformative" or "low confidence" studies that are generally not considered further in the IRIS assessment. The text anticipating this apparent exclusion of certain studies from future consideration based on their study evaluation determination (p. 6-6 of the handbook) does not represent a practicable, repeatable process. The text of the handbook addresses which confidence judgments correspond to more or less weight in later evidence synthesis and integration, acknowledging that such a determination may depend on what other studies on the exposure/endpoint are available and allowing for additional consideration on the basis of whether a confidence judgment results from sensitivity or other concerns. This handbook text gives the impression that certain studies are excluded from future review in the IRIS assessment process based solely on the individual study evaluation rating, when in practice such determinations are reached at other points in the process. During the presentation by Thayer (2021) and in written responses to the committee's questions, EPA clarified that "deficient" and "critically deficient" judgments can be quite varied and that "low confidence" studies are not typically excluded. However, EPA provided data from recent IRIS assessments showing that the proportion of human studies rated as "uninformative" and excluded from further consideration ranged from 0 to 50 percent, and 0 to 41.5 percent for animal studies. Thus, depending on the IRIS assessment, excluding studies at the study evaluation stage could lead to a substantial proportion of excluded studies due to a critically deficient rating in one domain. The importance of robust and transparent information, properly contextualized within the framework of the entire systematic review, suggests, as recommended in NASEM (2021), that study evaluation ratings should not be used to exclude studies. "Organizing the hazard review" or "evidence synthesis" are two possible process points for a decision about study

exclusion, with the exclusion or practical allocation of zero or nonzero weight to any given study being determined in the context of information available across studies.

Risk of Bias and Other Quality Assessment Items (Sensitivity and Reporting Quality)

Standard procedures for systematic review, which have a longer history in human clinical studies than in animal toxicological studies, evaluate studies on the basis of risk of bias and internal validity (Higgins and Thomas, 2019). The handbook augments study evaluation to consider, in both human and animal studies, the additional quality assessment items of sensitivity and reporting quality. The inclusion of these additional quality assessment items was anticipated in NRC (2014) and has precedent in animal toxicological studies (Hooijmans et al., 2018; NTP, 2019). However, the distinctions among the concepts of risk of bias, reporting quality, and sensitivity are not always clearly delineated in the handbook, and some considerations have clear overlap with more established concepts in systematic review. This mixing of concepts is prone to inhibit transparency and reproducibility in documenting how each concept is incorporated into domain judgments and study confidence ratings. Operationally, these concepts are sometimes considered collectively when evaluating domains, complicating an ultimate evaluation where elements of one concept can be incorporated into a judgment in multiple domains and be simultaneously cited as their own evaluation concerns. As such, it is difficult to see exactly how the procedure in the handbook would avoid "double counting" a single issue in multiple domains. Some concepts could conceivably be considered outside the study evaluation portion of the systematic review. Differences between the description of these concepts in the human epidemiological studies and the animal toxicological studies suggests two distinct study evaluation processes, yet both use the same terminology which furthers the confusion.

Risk of Bias

Of the three concepts, risk of bias and its attendant evaluation procedures show the most commonality and a high degree of concordance with other documented processes for systematic review. Reliance on the ROBINS framework for human studies has a clear precedent, although the committee notes there are documented difficulties in applying this framework in practice (Minozzi, et al., 2019; Igelström et al., 2021). The handbook's adaptations to address the types of studies typically encountered in environmental and occupational epidemiology are appropriate. IRIS program staff members are actively engaged in the development of a Risk of Bias in Non-randomized Studies of Exposures (ROBINS-E) (Morgan et al., 2017) counterpart to the existing ROBINS-I, as evidenced in the adaptations that appear in the handbook. Participation in this ongoing development renders IRIS staff well positioned to implement any eventual documented ROBINS-E framework, but the evolving nature of this framework is noteworthy, and limitations to the ROBINS-E framework have been noted (Bero et al., 2018).

Sensitivity

The handbook discusses the purportedly distinct concept of sensitivity as the ability of a study to detect a true association, with insensitive studies prone to produce "false negative" results. Low sensitivity in both the animal toxicological and human epidemiological studies is generally described as "bias toward the null," which might arise due to certain types of exposure

measurement error, and also in terms that relate directly to study precision, such as sample size, which would represent an issue distinct from bias. The committee viewed this notion of sensitivity as a less established construct than others used in systematic review, noting that it actually touches on several more established concepts, including (but not limited to) precision, risk of bias, generalizability, and other study features that relate to design or choice of endpoints that might be evaluated relative to refined population, exposure, comparator, outcome (PECO) statements (Higgins and Thomas, 2019). The *Cochrane Handbook* notes, "Bias should not be confused with imprecision. Bias refers to systematic error, meaning that multiple replications of the same study would reach the wrong answer on average. Imprecision refers to random error, meaning that multiple replications of the same study will produce different effect estimates because of sampling variation, but would give the right answer on average" (Higgins and Thomas, 2019, p. 179). Precision is related to the sample size of the study and confidence interval around the effect estimate, and a study that lacks precision may produce a "false negative" by virtue of random (sampling) variability. Thus, a null result due to lack of precision should not be described as "bias toward the null" as though it were a study feature (e.g., exposure measurement error) that might produce systematic error toward the null.

The *Cochrane Handbook* also notes, "Bias should not be confused with the external validity of a study, that is, the extent to which the results of a study can be generalized to other populations and settings" (Higgins and Thomas, 2019, p. 179). A study judged as insensitive due to the choice of population features (or exposure levels or follow-up periods) could produce what the handbook describes as a "false negative" by virtue of a lack of generalizability between the results of the study and the population (or exposure levels or follow-up periods) for which the risk assessment is of primary interest. That is, characteristics such as populations (or exposure levels or follow-up periods) dictate the precise exposure-response association that a particular study may address, and key questions about how that particular exposure-response association relates to the goal of the risk assessment should not be confused with "bias toward the null" or study quality, more generally. For example, EPA may posit a critical age at exposure, susceptible subgroup, exposure regimen of primary concern, window of exposure, and endpoints of interest. The systematic review may then identify a set of studies that address the agent of interest and outcome of interest, but classify a study as "low sensitivity" if it does not evaluate the ages, subgroups, exposure patterns, or temporal periods of effect that EPA has hypothesized are of most interest. Such a study may not be biased, and it may be adequately powered to detect the particular exposure-response association under study; yet EPA may still characterize it as "low sensitivity." In such cases, EPA has implicitly changed the question being evaluated—from one that is general ("Does the agent affect the endpoint?") to one that is narrower ("Does this study evaluate the effect of the agent on the endpoint in the period of follow-up, when we posit the effect will be largest?"). Judging a study as "low sensitivity" due to these types of study features should not downgrade the study's evaluation with regard to hazard identification but rather should be judged as addressing the agent's impact under a defined exposure scenario for a defined period of observation. A study designated as "low sensitivity" may provide useful evidence of effect heterogeneity (across different populations or exposure levels) or offer supporting evidence for other studies that consider more relevant exposure duration or follow-up periods (similar to the concept of a "negative control"). How exactly a particular exposure-response association being evaluated in a given study relates to the goals of the risk assessment could be captured through the study's relationship with the PECO statement, provided that such a PECO statement had been refined from the original broad PECO statement posited in the

problem formulation stage. Insofar as some of the human epidemiological or animal toxicological studies may be more or less relevant for the IRIS assessment, a general judgment of "whether there are factors in the design and conduct of the study that may reduce its ability to an observe an effect, if present" (EPA, 2020a, Table 6-1, p. 6-2) may not fully capture the different ways in which studies that target varying components of an exposure-response association might contribute to hazard identification. The systematic review infrastructure offers the opportunity to consider such studies in relation to the broad and refined PECO statements when judging relevance for the risk assessment, rather than use their features to determine sensitivity or study quality.

For evaluation of human epidemiological studies, the handbook briefly notes the concept of defining "an 'ideal' design (i.e., a study design with no risk of bias and high sensitivity) for the review question" (EPA, 2020a, p. 6-9) to be used as a reference point in gauging domain-specific ratings. The committee acknowledges the potential controversy surrounding the use of the "ideal design" in observational epidemiology when a randomized controlled trial is designated as the ideal. However, the committee emphasizes that while this concept is most typically used to assess risk of bias, it also encompasses (as noted on p. 6-9 of the handbook) study features, such as population definition, exposure window, and follow-up period. As such, the concept of an "ideal design" could potentially play a more explicit role in the handbook for clarifying how the particular exposure-response association being investigated in a given study relates to the "ideal" exposure-response association defined by a particular population of interest, exposure window, follow-up period, or other considerations.

Finally, some aspects of sensitivity discussed in the animal toxicological studies were viewed as overlapping with notions of reporting quality, possibly due to their inherently subjective nature. For example, considerations described as "outcome measures and results display" within the sensitivity domain include prompting questions such as "Does the level of detail allow for an informed interpretation of the results" and listed general considerations include "use of unreliable methods," which would both point toward considerations that could be considered in the reporting quality evaluation. Clear, concrete criteria and guidelines for each domain of study sensitivity would emphasize delineation with other study quality concepts and improve transparency and operationalization. Given that it only takes a single "critically deficient" element to be deemed "uninformative," the criteria for what constitutes a "critical deficiency" need to be unambiguously stated.

While many of the components of study sensitivity discussed in the handbook are important for study evaluation, the lack of alignment with other more established concepts in systematic review inhibits transparency of the study evaluation. While the handbook is clear that study features concerning sensitivity that may be evaluated in other bias domains should not be "double counted" in the sensitivity domain, a lack of guidance on how to allocate judgments about these features across multiple domains hinders operationalization and repeatability of the study evaluation process. A more focused description of the elements of sensitivity that are not overlapping with other parts of the systematic review process or other more established systematic review concepts would aid operationalization of the handbook. In addition to consideration of sensitivity relative to concepts such as precision and generalizability, concepts such as "directness and applicability" (NTP, 2019), judgments of a study's relation to PECO, or more concrete descriptions of the notion of an "ideal design" in the ROBINS framework (for the human epidemiological studies) would provide a more transparent way to document how a study may or may not contribute to ultimate risk assessment.

Overall, the handbook's discussion of sensitivity addresses many concepts that are important for study evaluation, but describing these distinct concepts with the term "sensitivity" is ambiguous and leads to a definition of sensitivity that is potentially overlapping with more established concepts of study evaluation and systematic review, as well the potential for some elements of study sensitivity to be evaluated relative to PECO statement(s). Consideration of these concepts is essential, but the ambiguity and overlap of concepts hinder operationalization, repeatability, and transparency of the study evaluation procedures. Issues relating to study precision and generalizability in particular might be appropriately separated from evaluation of study quality or bias. The committee acknowledges that there may be aspects of what is currently described as study sensitivity that may not be naturally captured in other systematic review concepts. Examples from Table 6-9 (p. 6-32) of the handbook that might warrant isolation as quality assessment items separate from risk of bias, precision, or generalizability include chemical administration and characterization and results presentation. If quality assessment items require isolation from other systematic review concepts, these aspects warrant a precise definition alongside an explicit procedure for operationalizing their consideration as distinct elements of study evaluation. References to such concepts in the literature on environmental health and other disciplines would strengthen the rationale for their use as distinct concepts in the systematic review context. Should EPA maintain study evaluation concepts that are currently described as sensitivity considerations and cannot be captured with more established terminology for systematic review, the use of a word other than "sensitivity" to describe these necessary aspects would avoid confusion with several other technical and colloquial meanings of the word "sensitivity."

Reporting Quality

The final quality assessment item considered in Chapter 6 of the handbook is reporting quality, which receives strikingly different discussion between the human epidemiological and animal toxicological studies. Inclusion of reporting quality as a quality assessment item is atypical in systematic review protocols but, as with sensitivity, has more recently emerged in systematic reviews of animal toxicological studies (SYRCLE [Hooijmans et al., 2018 and NTP, 2019). In the human studies, reporting quality is to be considered alongside risk of bias in each domain, whereas in the animal studies reporting quality is explicitly identified as a separate domain. The detail around reporting quality in animal toxicological studies is not mirrored in the discussion of human epidemiology studies. The committee appreciates the potentially broader set of factors that may dictate reporting quality in epidemiological versus experimental toxicological studies, but it questions the apparent guidance that notions of reporting quality require such different treatment in the human epidemiological versus animal toxicological studies. If there are specific reasons why reporting quality requires such a different role in these different types of studies, they are not made explicit in the handbook. The handbook does not provide sufficient guidance as to how reporting quality should be used to inform judgments in the risk of bias domains in the human studies, and the apparent subjectivity in this assessment inhibits transparency and operationalization.

The handbook text on experimental animal toxicology studies indicates that the IRIS program considers reporting quality to be a first consideration of whether the study has reported sufficient details to conduct a risk of bias and sensitivity assessment, noting that studies that do not report basic information are typically rated as "uninformative." The criteria for evaluating

reporting quality are unambiguous and what would constitute a "critically deficient" judgment in the reporting quality domain is clear, logical, and consistent with prior National Academies reports (Table 6-9 of the handbook, p. 6-32). As with some components of study sensitivity, the committee identifies opportunities for elements of reporting quality to have been considered at other points in the systematic review, for example, in relation to the PECO statement, where an inability to discern (e.g., due to poor reporting quality) the animal species, sex, or exposure route or duration might lead to exclusion of a study for an inability to judge its fitness with the PECO statement, much less evaluate the study for risk assessment. In addition, and also in common with the sensitivity quality assessment, the committee identifies what could be considered overlap between some of the cited features of reporting quality and other considerations that are more commonly regarded as risk of bias domains. For example, the risk of bias assessment could conceivably be expanded to include domains for "outcome ascertainment" and "participant selection" (as is done in the human epidemiological studies), in which case issues related to reporting on the assays or procedures used to measure the endpoint or the species or sex of the animal could be naturally subsumed within a more standard risk of bias assessment.

Overall, the committee notes the usefulness of the specific reporting quality elements described in the critical information for animal toxicological studies, but it could not identify why a similar level of detail was absent from the discussion of human studies or why some concepts such as test animal and exposure methods required consideration outside of traditional risk of bias domains or PECO considerations. A transparent and repeatable review process requires either explicit guidance for how to incorporate reporting quality into individual risk of bias domains or an explicit rationale for two separate processes for assessing reporting quality in human and animal study evaluation. In either case, explicit rationale for isolating elements of reporting quality from established systematic review concepts would aid operationalization of the study evaluation process.

Co-exposures

The committee recognizes the difficulty and evolving landscape of methodology for handling co-exposures in environmental and occupational epidemiology and animal toxicology, which might include chemical mixtures or non-chemical stressors. The handbook confines this discussion to the confounding domain in human epidemiological studies, but broader considerations are evolving in the literature. Consequences of co-exposures for risk assessment could include (but not be limited to) (1) multicolinearity and confounding of the health effect associated with the primary exposure, synergy, or interaction such that the health effect of one agent may be altered by the presence of a co-exposure; and (2) generalizability, where changes in co-exposures across populations hinder the transportability of results across populations. The near ubiquity of complex exposure mixtures warrants future consideration in refinement of study evaluation processes for IRIS assessment as this methodological area continues to evolve.

Study Evaluation Methods for PK Models, PBPK Models, and Mechanistic Evidence

Section 6.5 of the handbook indicates that a PK model or PBPK model must be evaluated before it can be accepted for use in an IRIS assessment. It appears that the evaluation would be carried out in parallel with, rather than as part of, a systematic review. Scientific and technical criteria for judging the suitability of a model include evaluating:

- Representation of a chemical mode of action by the model's structure and equations, based on available scientific information,
- Availability of the computer code and apparent completeness of parameter listing and documentation,
- Implementation of the conceptual model in the computational code,
- Use of appropriate parameters in the model, and
- Reproducibility of model results reported in scientific publications.

The committee considers the model evaluation approach described in the handbook to be adequate.

Mechanistic Evidence

The pilot testing approach for mechanistic data is described in Section 6.6 of the handbook. The committee considers that approach to be reasonable, as it is analogous to the approach for evaluation of animal toxicological studies (see handbook Table 6-10, beginning on p. 6-48). The issues identified with respect to sensitivity in animal studies are also relevant to the methods for evaluation of mechanistic evidence, and need to be considered accordingly.

FINDINGS AND RECOMMENDATIONS

Findings and Tier 1 Recommendations

Finding: The handbook describes circumstances under which a study may be excluded from the systematic review based on the outcome of the study evaluation. However, such exclusion is inconsistent with recent recommendations to incorporate study evaluation ratings within the context of evidence synthesis. For example, see the 2021 National Academies report *The Use of Systematic Review in EPA's Toxic Substances Control Act Risk Evaluations* and the *Cochrane Handbook for Systematic Reviews of Interventions*.

> **Recommendation 4.1:** The handbook should not use the results of study evaluation as eligibility criteria for the systematic review. *[Tier 1]*

Finding: The quality assessment item described as "sensitivity" covers important concepts, but is ambiguous and under-operationalized, as it covers aspects of internal validity, external validity, and statistical precision that overlap with other more commonly accepted features of systematic review that may be better assessed at other stages of the systematic review process.

> **Recommendation 4.2:** EPA should evaluate whether aspects currently captured in the notion of "sensitivity" might be better described in the handbook with more established terminology (e.g., precision or generalizability) or better addressed at other points of the systematic review (e.g., risk of bias assessment or evaluation relative to PECO statement[s]). Otherwise, the handbook should provide a more concrete definition of "sensitivity" and a procedure for operationalizing its use in the study evaluation step. *[Tier 1]*

Finding: The use of reporting quality as a distinct quality assessment item for study evaluation is not standard for systematic reviews, and procedures for evaluating reporting quality are very different for the human epidemiological and animal toxicological studies. Reporting quality is included within other evaluation domains for human studies, with almost no specific guidance on how to incorporate reporting quality within these domains. This presents the possibility of downgrading a study quality rating without any specific evaluation criteria. For animal studies, reporting quality is described in detail as a separate evaluation domain but with some aspects that overlap with other systematic review concepts such as risk of bias or external validity. The handbook notes that reviewers should reach out to study authors to obtain missing information in animal toxicological studies, but it provides no such reference to obtaining information from authors of human epidemiological studies.

> **Recommendation 4.3:** The handbook should address the apparent difference in assessing reporting quality between the human epidemiological and animal toxicological studies by either (1) assessing reporting quality similarly in both types of studies or (2) providing an explicit rationale for why the concepts require different assessment procedures in different types of studies. In either case, the handbook should provide an explicit rationale for isolating elements of reporting quality from established systematic review concepts and evaluate whether aspects currently described as reporting quality might be better addressed at other points of the systematic review process. *[Tier 1]*

Finding and Tier 2 Recommendation

Finding: Operationalizing the steps for conducting the study evaluation is opaque from the text of the handbook but greatly clarified by presentation of how many steps are operationalized with the Health Assessment Workspace Collaborative (HAWC).

> **Recommendation 4.4:** EPA should redraft the handbook to harmonize descriptive text with illustrations of how steps for conducting the study evaluation are operationalized in practice, using the presented material as needed. It is evident that all IRIS assessments will be performed using HAWC or similar software. Screen shots illustrating key steps would greatly improve conceptual understanding and operationalization (similar to Figures 6-2 and 6-3 of the handbook). Additional step-by-step instructions with screenshots could be included as supplementary material or links to other content. *[Tier 2]*

Finding and Tier 3 Recommendation

Finding: Consideration of how multiple co-exposures or chemical mixtures may relate to assessing risk of bias or other quality assessment items is limited to a discussion of potential confounding, which may not fully capture the possible impact on systematic review.

> **Recommendation 4.5:** EPA should monitor ongoing methodological development in assessing risk amid environmental co-exposures or mixtures in order to update explicit guidance on their potential roles in the evaluation of human epidemiological and animal toxicological studies. *[Tier 3]*

5

Evidence Synthesis

In this chapter, the committee reviews the following chapters of the *ORD Staff Handbook for Developing IRIS Assessments* (the handbook): "Analysis and Synthesis of Human and Experimental Animal Data" (Chapter 9), "Analysis and Synthesis of Mechanistic Information" (Chapter 10), and the portions of "Evidence Integration" (Chapter 11) that discuss synthesis (EPA, 2020a). The committee's review of those chapters considered whether the approaches to evidence synthesis are scientifically sound. It also examined whether the considerations presented in those handbook chapters are sufficiently broad to allow for application to the wide range of scenarios expected to be encountered when applied to individual Integrated Risk Information System (IRIS) assessments. In addition, the committee considered whether the methods presented in those handbook chapters are sufficiently clear in describing the intent to synthesize the relevant evidence and incorporate study evaluation conclusions, regardless of the study results (Question 6a, Appendix B).

OVERVIEW OF EVIDENCE SYNTHESIS AS COVERED IN THE HANDBOOK

The U.S. Environmental Protection Agency (EPA) IRIS process for evidence synthesis is presented in Chapters 9, 10, and 11. The purpose of these chapters is to summarize and interpret the results across all informative health effect studies within both human and animal streams and to integrate those data to draft hazard synthesis sections describing human and animal toxicity data. Approaches for analyzing mechanistic data on endpoints that lead to, or modify, the development of adverse outcomes are provided in Chapter 10. In addition to what was provided in the handbook, the committee relied significantly on EPA's presentation (Thayer, 2021). Screenshots of the Health Assessment Workspace Collaborative (HAWC) within-evidence stream judgment process were critical for understanding the approaches incompletely described in Chapter 9 and then redundantly, though not consistently, reiterated in Chapter 11.

A goal of the handbook is to enhance transparency of EPA's "guided expert-judgment" process. This is consistent with recommendations from the 2014 National Academies report *Review of EPA's Integrated Risk Information System (IRIS) Process* to

> develop templates for structured narrative justifications of the evidence-integration process and conclusion. The premises and structure of the argument for or against a chemical's posing a hazard should be made as explicit as possible, should be connected explicitly to evidence tables produced in previous stages of the IRIS process, and should consider all lines of evidence (human, animal, and mechanistic) used to reach major conclusions. (NRC, 2014, p.106)

CRITIQUE ON EVIDENCE SYNTHESIS

The HAWC approach presented by EPA (Thayer, 2021), including factors that increase certainty and factors that decrease certainty, seems to be a meaningful move toward transparency. That said, Table 9-1 (p. 9-3) and Table 11-1 (p. 11-5) are inconsistent in descriptions of the considerations, and Table 11-1 appears more representative of the process. Chapters 9 and 11 are not clearly organized with respect to where synthesis stops and integration starts. As there is no conclusion at the end of Chapter 9, it is not clear what is accomplished in that chapter. If two chapters would cover synthesis/integration, the first chapter could address Step 1 (pp. 11-2 to 11-3; within-stream judgments) and the second chapter could address Step 2 (across streams).

Unit of Analysis

Given the committee's understanding of the handbook, the process or method for evidence synthesis, particularly in Chapter 9, is not clearly or consistently described. With the understanding that HAWC is a dynamic system, Table 9-1 is not consistent with the column headers and criteria in the HAWC interface, namely Studies, outcomes, and confidence; Summary of key findings; Factors that increase certainty (consistency, dose-response, coherence of effects, larger or concerning magnitude of effect, mechanistic evidence providing plausibility, medium or high confidence studies, other); and Factors that decrease certainty (unexplained inconsistency, imprecision, lack of expected coherence, evidence demonstrating implausibility, low confidence studies). Furthermore, Table 9-2 (p. 9-6) introduces individual and social factors that may increase susceptibility to exposure-related health effects but without specifying how those factors are considered in the within-stream evidence judgment. Presumably in the "Factors that increase/decrease certainty" column both susceptibility and publication bias can be noted. These options could be more clearly categorized as "Other" if Table 9-1 was redesigned to match the HAWC judgment boxes. Also, the role of natural experiments in Chapter 9 is out of step with how risk of bias is discussed in the study evaluation chapter. Natural experiments are a study design, which can be evaluated using risk of bias criteria suited to that study design; they are not a consideration for evidence synthesis as stated in Table 9-1. In general, using HAWC, the methods associated with within-stream evidence judgments can be implemented in a transparent manner, but that is not yet clearly communicated in the handbook.

The evidence synthesis approach in general is sufficiently broad to allow for application to a wide range of scenarios. Yet, the committee often stumbled on different terminology used and that supported the need for glossary definitions for "endpoint," "[health] outcome," "health effect," "bodies of evidence," and "streams." In a systematic review framework, the individual "bodies of evidence" need to be prespecified. It is recommended in Chapter 3 of this report that this be accomplished through the development of a series of "refined" population, exposure, comparator, outcome (PECO) statements informed by the literature inventory/evidence map (which utilized a "broad" PECO statement) and specified in the assessment protocol.

A related element of confusion is the "unit of analysis" of the synthesis step. In Chapter 9, it appears that there are only two "bodies of evidence"—animal and human—for each health effect. On the other hand, in Chapter 11, it is noted that "inferences" may be drawn from "a finer level of specificity of effect" and then used "to draw conclusions about the broader health effect categories." Additionally, Chapter 11 states that "evaluation of strength of … evidence will

preferably occur at the most specific health outcome level possible" (EPA, 2020a, p. 11-8). Indeed, in Table 11-1, it appears that the template evidence profile table suggests that "separate rows by outcome," possibly also by study design, are permitted and that each "row" that undergoes synthesis results in a strength of evidence judgment. However, in EPA's presentation, an example was presented where liver effects in animals were separated by "clinical chemistry" and "histopathology," but only a single strength of evidence judgment was presented (Thayer, 2021).

A corollary issue is that in a systematic review framework, mechanistic and toxicokinetic (TK) data or pharmacokinetic (PK) and physiologically based pharmacokinetic (PBPK) models play a narrow role in evidence synthesis of animal and human bodies of evidence, primarily limited to the evaluation of consistency. However, as described in Chapter 3 of this report, in some cases, mechanistic endpoints may be considered as discrete endpoints that will undergo evidence synthesis. If those endpoints involve in vivo animal or human data, the approach to evidence synthesis for animal and human data streams is appropriate.

In vitro or non-traditional evidence streams present more of a challenge and are currently an area of rapid evolution. The handbook has outlined a reasonable approach to evaluating in vitro data in Chapter 10, with the synthesis approach shown in Chapter 11. These represent reasonable approaches but are likely to change substantially with time.

Choices of Methods

The evidence synthesis approach outlined in the handbook appears to be a "hybrid" of a guided expert judgment approach (e.g., Hill, 1965; IARC, 2019; Samet et al., 2020) and a more structured approach (e.g., Grading of Recommendations Assessment, Development and Evaluation [GRADE] [Balshem et al., 2011; Guyatt et al., 2011]; NTP, 2019). For instance, while the considerations delineated appear primarily to be based on the Bradford Hill considerations (with the addition of "study confidence," which includes risk of bias), the structure of the analysis resembles GRADE but without the explicit up-rating and down-rating of evidence. The committee reiterates the view of the NRC (2014) report withholding an explicit recommendation to choose either guided expert judgment or a structured approach. The committee notes that the handbook's evidence synthesis framework does not rely on the Bradford Hill considerations (1965) alone but also is aligned with multiple more modern evidence evaluation frameworks, including Cochrane (Higgins and Thomas, 2019), GRADE, NTP (2019), and the Navigation Guide (Woodruff and Sutton, 2014.

More problematically, however, the considerations outlined in evidence synthesis appear to mix in some elements of evidence integration, particularly through the concepts of coherence and biological plausibility. The result is confusing, especially when combined with the challenging transition between Chapter 9 and 11, described above. Most systematic review frameworks do not include the concept of "coherence" in the evidence synthesis step, because that would involve bringing in information from outside the evidence being evaluated. Moreover, the application of "coherence" as described in the handbook appears more appropriate either (1) during planning of the assessment (the biological relationship among different endpoints) or (2) during evidence integration (through incorporation of mechanistic data). The concept of "biological plausibility" is largely absent in most systematic review frameworks except perhaps in the concept of "indirectness" used by GRADE and Cochrane. Overall, the applications of "biological plausibility" in the handbook appear to either (1) address considerations already

covered elsewhere, such as consistency; (2) be more appropriate to consider during evidence integration; or (3) involve comparison with data on mechanistic effects. The latter two are more appropriate to incorporate during the evidence integration step.

Finally, the consideration of natural experiments is inappropriate in Chapter 9 of the handbook. They should already have been considered in "study confidence," since the criteria used for study evaluation will depend on the design of the study and these criteria are discussed in Chapter 4 of this report.

The remaining considerations—study confidence, consistency, strength (effect magnitude) and precision, and biological gradient/dose-response—are closely aligned with existing systematic evidence evaluation frameworks, such as Cochrane, GRADE, NTP, and the Navigation Guide. This is appropriate because the Bradford Hill considerations encompass both evidence synthesis and evidence integration, while modern evidence evaluation frameworks make an explicit distinction between them in order to promote greater transparency and consistency.

Logistics

Chapter 9 of the handbook provides a limited discussion on data visualization (p. 9-7) including a forest plot for trichloroethylene, but there is much less emphasis on data visualizations within evidence streams compared to that in Chapter 6. Trend analyses could also be helpful in the analysis and synthesis of within-stream studies.

The committee noted a concern that the IRIS process could become very bogged down during the within-stream judgment process, given the statement in Chapter 9 that "evidence synthesis is an iterative process" (EPA, 2020a, p. 9-1). This concern is compounded by the lack of clarity around who is actually doing the analysis ("assessment team and disciplinary workgroups (as needed)"). It would be useful to more clearly define who is responsible for conducting the synthesis and indicate for quality control that some consensus is needed on the final judgments. Additionally, Table 9-1 suggests in "coherence" that "these analyses may require additional literature search strategies." It is not clear to the committee why additional literature review would arise at this late stage of the assessment. Collectively, these concerns reflect the committee's concern that the timeliness of evaluations could be adversely impacted based at least on how Chapter 9 of the handbook is presented.

FINDINGS AND RECOMMENDATIONS

Findings and Tier 1 Recommendations

Finding: The handbook is confusing as to the transition from the synthesis step (within a data stream) to integration (across data streams), in particular because many considerations for synthesis are repeated (with slight variation) in Chapters 9 and 11.

> **Recommendation 5.1:** The handbook should consolidate its discussion of evidence synthesis in a single place. The discussion should include all of the considerations involved in making strength of evidence conclusions (currently in Chapter 9), as well as the criteria for different strength of evidence judgments (currently in Chapter 11). The handbook chapter describing synthesis of evidence should end with the methods for

reaching strength of evidence conclusions for each unit of analysis, and how these are carried forward to evidence integration. *[Tier 1]*

Finding: The "unit of analysis" for evidence synthesis and the strength of evidence conclusion is unclear, in terms of the breadth or narrowness of the evidence being synthesized. Although the handbook states that the evaluation of strength of evidence "will preferably occur at the most specific health outcome possible," the process for doing so is not clear (EPA, 2020a, p. 11-8). For instance, in a case where liver clinical chemistry and liver histopathology (both in animals) are synthesized separately, will they *each* receive a strength of evidence judgment, will there be a *single* overall strength of evidence judgment for liver toxicity, or will strength of evidence judgments occur *both* individually and together?

Recommendation 5.2: The unit of analysis for evidence synthesis and strength of evidence conclusions should be clearly defined as specified by the refined PECO statements recommended in Chapter 3 of this report. For example, a unit of analysis could be defined at the endpoint level (e.g., clinical chemistry) or outcome level (e.g., liver toxicity). If judgments may be made at both the endpoint and health outcome levels, details should be provided on how these judgments and the methods used to make them are distinct from each other. *[Tier 1]*

Finding: The guided expert judgment approach that EPA uses does not describe how to specify the initial rating for the strength of evidence evaluation. The approach does not include explicit criteria for up-rating and down-rating evidence and does not include sufficient information to operationalize the approach.

Recommendation 5.3: The handbook should provide justification for the initial rating for strength of evidence, as well as more detailed operationalization of the criteria used to upgrade or downgrade the evidence. *[Tier 1]*

Finding: The considerations around mechanistic and TK data and PK or PBPK models outlined in evidence synthesis appear to mix in some elements of evidence integration, particularly through the concepts of coherence and biological plausibility, because they rely heavily on evidence outside of the "unit of analysis."

Recommendation 5.4: The handbook should restrict the applications of mechanistic and TK data or PK models in evidence synthesis and strength of evidence judgments to those relevant to each individual unit of analysis, such as addressing consistency and indirectness of evidence. Other applications of mechanistic and TK data or PK models, such as addressing coherence and elements of biological plausibility, could be addressed in evidence integration, either as a separate evidence stream or as support for the human or animal evidence streams. This recommendation should be implemented in the planning stage and reflected in the protocol (see Recommendation 3.9). *[Tier 1]*

Finding and Tier 2 Recommendation

Finding: Consideration of "natural experiments" is not part of evidence synthesis but rather part of study evaluation.

> **Recommendation 5.5:** EPA should remove discussion of natural experiments in the context of evidence synthesis in the handbook. That discussion could possibly be repurposed in the context of the handbook's Chapter 6 "Study Evaluation." *[Tier 2]*

No Tier 3 recommendations were presented in this chapter.

6

Evidence Integration

This chapter provides the committee's review of Chapter 11 of the *ORD Staff Handbook for Developing IRIS Assessments* (the handbook), "Evidence Integration" (EPA, 2020a). The committee considered whether the approaches described in that chapter of the handbook are scientifically sound and appropriate for integrating the various types of evidence relevant to investigating the potential for human health effects from exposure to environmental chemicals. The committee also commented on approaches for developing overall integration judgments that involve using three versus five categories of judgment levels (Questions 6b and 7 in Appendix B).

OVERVIEW OF THE HANDBOOK'S MATERIAL ON EVIDENCE INTEGRATION

Chapter 11 of the handbook maps out the process the Integrated Risk Information System (IRIS) program intends to use for integrating human, animal, and mechanistic data to assess the human health hazard of the chemical being reviewed. Details are provided on the proposed evidence integration narratives with summary judgments and supporting rationale for these decisions. Although evidence integration is typically a stand-alone step in the systematic review process used to summarize data across evidence streams, there is overlap in the handbook with Chapter 9 "Analysis and Synthesis of Human and Experimental Animal Data," which provides general considerations for evidence synthesis within a data stream.

The National Institute of Environmental Health Science's Office of Health Assessment and Translation (OHAT) *Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration* describes three distinct steps following the risk of bias assessment: synthesizing the evidence (within human and animal streams), rating the confidence in the body of the evidence, and integrating the evidence (integrating human and animal evidence and considering mechanistic evidence) (NTP, 2019). The judgments developed in the evidence integration step are intended to directly inform the hazard identification and dose-response analysis. These steps, which OHAT defines as sequential but independent steps in the systematic review process, are nearly indistinguishable in Chapter 11 of the handbook.

RESPONSIVENESS TO PREVIOUS NATIONAL ACADEMIES REPORTS

Evidence integration has long been a challenge in toxicological systematic reviews, and previous National Academies reports have provided the U.S. Environmental Protection Agency (EPA) with several suggestions to improve the process. The 2011 National Academies report *Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde* recommended the development of a standardized approach for evidence integration (NRC, 2011). The 2014 National Academies report *Review of EPA's Integrated Risk Information System (IRIS) Process* commended EPA on its progress (NRC, 2014). That report provided

additional recommendations to the IRIS program for developing a more systematic process for evidence synthesis and integration to enhance transparency, efficiency, and scientific validity. The development of standardized, structured evidence tables was recommended by both the 2011 and 2014 reports to support the evidence judgments and narratives. The 2018 National Academies report *Progress Toward Transforming the Integrated Risk Information System (IRIS) Program: A 2018 Evaluation* assessed changes that EPA implemented or planned to implement in response to the recommendations in the 2014 National Academies report (NASEM, 2018). Based on a review of materials presented by EPA to the authoring committee, the 2018 report indicated that the process and framework for evidence integration were consistent with state-of-the-art approaches taken by other scientific agencies, such as the National Toxicology Program, that face similar challenges.

The use and incorporation of mechanistic data in systematic reviews in toxicology continue to be a challenge, not only for EPA but for the systematic review community in general. The 2014 National Academies report acknowledged the problematic use of various kinds of mechanistic data, specifically where in the process it should be incorporated and how to do so systematically. The 2014 report noted that solid conclusions about causality can be drawn without mechanistic information, for example, when there is strong and consistent evidence from animal or epidemiology studies, suggesting that mechanistic types of data are not necessarily needed in the integration step in order to make a judgment of the evidence.

CRITIQUE OF METHODS FOR EVIDENCE INTEGRATION

There are many strengths to Chapter 11 of the handbook as it is written now. Although the organization of this chapter has some issues, the approaches to strength of evidence judgments are very detailed and well described, and are likely to add to the transparency and consistency of the IRIS process for developing assessments. Chapter 11 also indicates it is preferable that at least two reviewers form evidence integration judgments independently. Involving multiple reviewers is an important feature of a systematic review, although it is unclear whether the reviewers remain the same throughout the review.

A tremendous amount of work has been done to improve evidence integration in risk assessment. However, there are still issues in the handbook that need to be addressed. The terms "synthesis," "integration," and "strength of the evidence" appear to be used almost interchangeably, when in fact these should be distinct steps in the systematic review process. As previously discussed, Chapter 11 overlaps somewhat with Chapter 9, and yet the chapters show inconsistencies in the considerations for evidence synthesis within a stream without a clear discussion as to the difference. For instance, Chapter 9 (Table 9-1 "Important considerations for evidence syntheses", p. 9-3) lists study confidence, consistency, strength (effect magnitude) and precision, biological gradient/dose-response, coherence, mechanistic evidence related to biological plausibility, and natural experiments. However, Chapter 11 (Table 11-2 "Considerations that inform evaluations and judgments of the strength of the evidence", p. 11-10) lists risk of bias/sensitivity, consistency, strength (effect magnitude) and precision, biological gradient/dose-response, coherence, and mechanistic evidence related to biological plausibility. While most of the terms are the same, the terms "study confidence," "natural experiments," "risk of bias," and "sensitivity" only appear in one or the other tables.

The handbook's discussion of supporting studies, including those providing mechanistic data, does not indicate a clear role throughout the assessment, particularly during the evidence

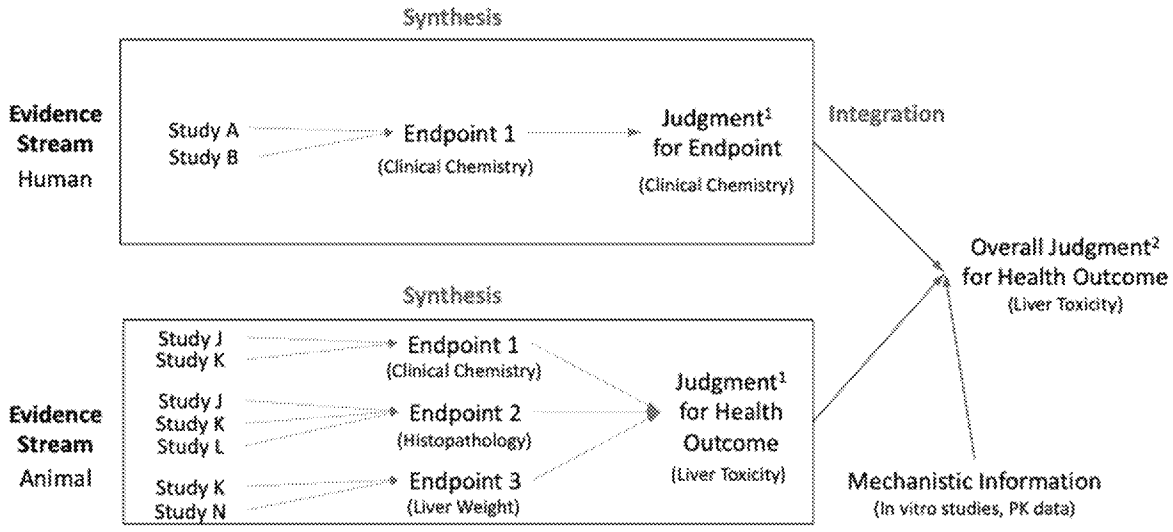
integration step. Mechanistic data appear to be used to support the strength of evidence conclusions of the individual data streams or when there is a lack of evidence in a single stream, and yet some aspects of the handbook treat mechanistic data as a separate stream. Table 11-1 of the handbook (p. 11-5) is an example of how these concepts have been intermingled, as it includes aspects of strength, synthesis, and the overall inferences across streams. Table 11-2 of the handbook (p. 11-10), appropriately titled "Considerations that inform evaluations and judgments of the strength of the evidence," focuses on evaluating the data streams individually and therefore does not follow the chapter description of integrating across streams. The steps outlined in Figure 11-1 of the handbook (p. 11-3) are also unclear, and the figure lacks information on how multiple evidence conclusions are narrowed down to a single judgment. The synthesis step is an analysis of "like" studies to form the basis of conclusions; all of these studies are to have gone through the same study quality evaluation, including risk of bias. The inclusion of toxicokinetic (TK) data, physiologically based pharmacokinetic (PBPK) models, and mechanistic data (outside of the precursor or mechanistic endpoints that are included in the human or animal "effects" studies) requires the use of very different approaches for evaluating the study and model quality (e.g., EPA's quality assurance project plan for PBPK models [EPA, 2018]). Furthermore, the role of TK data, PBPK models, and mechanistic data is to answer questions that arose in the synthesis step. (e.g., If toxicity was seen in the oral gavage study but not the drinking water study, are the findings consistent or not when accounting for TK considerations across exposure routes?)

Clarification of the evidence synthesis and integration processes also needs to address expectations for dose-response, as described in Chapter 12:

> Ideally, the hazard synthesis and integration has clarified any important considerations, including mechanistic understanding, that would indicate the use of particular dose-response models, including chemical-specific or biologically based models, over more generic models (see Chapter 13). These considerations also include whether linked health effects within and between organ systems should be characterized together, as well as whether there is suitable mechanistic information to support combining related outcomes or to identify internal dose measures that may differ among outcomes (generally for animal studies). (EPA, 2020a, p. 12-2, lines 10-16)

While aspects such as linking of multiple health effects or evaluating whether animal toxicological outcomes are species-specific may be necessary for hazard identification, it is likely that mechanistic and TK considerations will not be required. When and how to consider these data (e.g., during synthesis and integration or subsequently) needs to be defined more broadly so that mechanistic and TK data are not neglected in the assessment process.

As with other chapters in the handbook, consistent use of terminology is critical throughout Chapter 11. Multiple terms are used both independently and interchangeably, including "body of evidence," "evidence stream," "endpoint," "outcome," and "health effect." This causes confusion in the terminology surrounding the unit of analysis and makes the description of the process difficult to interpret and nearly impossible to follow. A terminology map, such as the example shown in Figure 6-1 below, would help to clarify the types of data being referred to, how the data move through the process, and where judgments will be applied.



[1] Judgments for health outcomes within an evidence stream include *robust, moderate, slight, indeterminate,* and *compelling evidence of no effect.*
[2] Judgments of health evidence across evidence streams include *evidence demonstrates, evidence indicates, evidence suggests but is not sufficient to infer, evidence inadequate,* and *strong evidence supports no effect.*

FIGURE 6-1 Example of a terminology map illustrating units of analysis.
Note: The figure shows the points at which synthesis, integration, and strength of evidence judgments may occur. In this example, the units of analysis for integration are the endpoint (clinical chemistry) for the human evidence stream, and the health outcome (liver toxicity) for the animal evidence stream.

EPA asked the committee to consider whether the approaches described in Chapter 11 are scientifically sound and appropriate for integrating the various types of evidence relevant to investigating the potential for human health effects from exposure to environmental chemicals.

As described in Chapter 5 of this report, the handbook is confusing as to the distinction between evidence synthesis and evidence integration, making it difficult to determine if the evidence integration process as presented in the handbook is scientifically sound and appropriate. In particular, there is inconsistency concerning whether the judgment as to a strength of evidence rating is (1) the direct result of synthesis at the endpoint level (as shown for clinical chemistry in slide 34 of Thayer [2021]) or (2) the result of synthesizing multiple outcomes (as shown for liver clinical chemistry and liver histopathology in slide 35, where the individual animal outcomes were synthesized separately and then given a single judgment). The addition of mechanistic evidence as a separate stream (Table 11-1 of the handbook (p. 11-5)) and as factors that may increase or decrease certainty in the animal and human streams (Thayer, 2021, slide 35) compounds the difficulty in interpreting this chapter.

In addition, EPA asked the committee to comment on approaches using five categories versus three categories for drawing evidence integration conclusions. The agency also asked which approach is recommended and why, and whether any specific refinements are needed.

Table 11-5 of the handbook (p. 11-22) presents five categories of evidence integration judgments for characterizing potential human health hazards regarding environmental exposure to a chemical:

- Evidence demonstrates that [*chemical*] causes [*health effect*].
- Evidence indicates that [*chemical*] likely causes [*health effect*].
- Evidence suggests that [*chemical*] may cause [*health effect*].
- Evidence is inadequate to assess whether [*chemical*] may cause [*health effect*].
- Strong evidence supports no effect.

The IRIS program included a three-category approach for evidence integration in several systematic review protocols proposed in EPA (2019a,b) and EPA (2020c):

- Sufficient evidence for hazard.
- Insufficient evidence.
- Sufficient evidence to judge that a hazard is unlikely.

There are potential advantages to either set of categories. For example, five categories would align the approach for noncancer judgments with the 2005 cancer guidelines (EPA, 2005), while three categories might reduce variation in judgments for different health outcomes and endpoints considered for a particular chemical and may make the process simpler to implement. However, the committee sees no strong scientific rationale for favoring one approach over the other. The committee has determined that a comparative assessment of the various factors related to using each framework in an IRIS assessment context and recommending one of them are important activities, but are outside of the committee's task of reviewing the handbook.

FINDINGS AND RECOMMENDATIONS

Finding: Chapter 11 of the handbook contains information on three sequential but independent steps in a typical systematic review process: synthesizing the evidence (units of analysis within human and animal streams), rating the confidence in the body of the evidence, and integrating the evidence (integrating human and animal evidence and considering mechanistic evidence).

> **Recommendation 6.1:** The handbook should separate and delineate the chapters with respect to evidence synthesis within a stream and evidence integration across data streams (see Recommendation 5.1). Synthesis, integration, and the judgments that are used to rate the evidence need to be clearly defined, distinct steps. *[Tier 1]*

Finding: The process outlined in Chapter 11 of the handbook remains unclear, with ambiguous language and a lack of stepwise instructions adding to the confusion. The process was more clearly described in Thayer (2021).

> **Recommendation 6.2:** EPA should supplement Chapter 11 of the handbook with additional figures and examples from other IRIS assessment documents. The addition of a terminology map (endpoint, outcome, synthesis, integration) would help define the steps that should occur at each level and, more specifically, what data are to be synthesized and where to expect the judgment narratives to be provided. *[Tier 1]*

Finding: The role of mechanistic data discussed in Chapter 11 (i.e., whether they are their own stream or simply used to support the human and animal streams) remains unclear. As noted by

the committee in Chapter 2, there may be multiple roles for mechanistic data in an IRIS assessment. While many of these are described in Chapter 11 of the handbook, it is not clear as to which of these roles constitute a separate stream (as shown in Table 11-1, p. 11-5) and which are simply used to support the human and/or animal streams (as described in Section 11.1 and Table 11-2 (p. 11-10) of the handbook.

> **Recommendation 6.3:** The handbook should clearly define the roles of mechanistic data and other supporting data in evidence integration and throughout the entire IRIS assessment development process (see Recommendations 3.9, 3.10, and 5.4). *[Tier 1]*

No Tier 2 or Tier 3 recommendations were presented in this chapter.

7

Hazard Considerations and Study Selection for Deriving Toxicity Values

In this chapter, the committee reviews Chapter 12 "Hazard Considerations and Study Selection for Deriving Toxicity Values" and Chapter 13 "Derivation of Toxicity Values" of the *ORD Staff Handbook for Developing IRIS Assessments* (the handbook) (EPA, 2020a). The committee considered whether those handbook chapters provide appropriate considerations for identifying data sets for dose-response analysis based on systematic review conclusions. The committee also considered whether the basic methods for dose-response modeling and deriving toxicity values are consistent with the state of the science and are presented in the handbook with sufficient clarity (Question 8 in Appendix B.)

OVERVIEW OF THE HANDBOOK'S CHAPTERS 12 AND 13

Chapter 12 of the handbook summarizes the considerations and approaches involved in arriving at hazard judgments to set priorities among health outcomes and select the studies to be used for toxicity value derivation. That chapter also outlines considerations to be used for study selection and information for combining dose-response data.

Chapter 13 discusses the processes used to derive various toxicity values, referencing additional U.S. Environmental Protection Agency (EPA) guidance documents as needed. The chapter also includes directions for characterizing sources of uncertainty and confidence in the derived toxicity values.

RESPONSIVENESS TO PREVIOUS NATIONAL ACADEMIES REPORTS

Previous reviews by the National Academies commented heavily on the processes used for selecting studies, and the current handbook attempted to address these concerns. The National Academies report *Review of EPA's Integrated Risk Information System (IRIS) Process* (NRC, 2014) recommended that EPA develop criteria for determining when evidence is sufficient to derive toxicity values. Pertinent to this recommendation, the handbook states that "for both cancer and noncancer hazards, preference is given to health effects (or outcomes) and cancer types with stronger evidence integration conclusions" (EPA, 2020a, p. 12-4). Chapter 11 of the handbook discusses the evidence integration judgments used for noncancer endpoints and cites EPA's *Guidelines for Carcinogen Risk Assessment* (EPA, 2005) to provide additional information for evaluating carcinogenic evidence.

The 2014 National Academies report also recommended that EPA "continue its shift toward the use of multiple studies rather than single studies for dose–response assessment" by developing formal methods for combining data from multiple studies (NRC, 2014, p. 129). The handbook includes a section that provides additional considerations and methods for evaluating and combining dose-response data from multiple studies as an alternate approach to using data

from a single study. The 2018 National Academies report *Progress Toward Transforming the Integrated Risk Information System (IRIS) Program: A 2018 Evaluation* cautioned that "care must be taken when combining results within or between studies in developing dose–response relationships inasmuch as multiple mechanisms, each with its own potential dose–response relationship, might be involved" (NASEM, 2018, p.11). As indicated in Chapter 13 of the handbook, methods for the derivation of toxicity values are detailed in several historical guidance documents and are largely limited to those previous documents. The 2014 National Academies report recommended that EPA develop IRIS-specific guidelines for communicating uncertainty, and, in response, Chapter 13 of the handbook includes a section on characterizing uncertainty and communicating confidence in the derived toxicity values.

CRITIQUE OF METHODS FOR STUDY SELECTION AND DERIVING TOXICITY VALUES

Chapters 12 and 13 of the handbook focus on steps to be taken after systematic review, specifically the processes by which studies are selected and toxicity values are derived, and many aspects of sections in those chapters positively add to the overall assessment process. Chapter 12 aims to standardize the selection of studies for toxicity value development by providing considerations and case-specific recommendations. Chapter 13 addresses development of toxicity values, with most of the sections referring to other EPA guidance documents. Several aspects of both Chapters 12 and 13 would benefit from clarification. Overall, although improvements are evident, the level of detail in these chapters is not sufficient to provide a step-by-step guide for those involved in developing IRIS assessments.

The committee considered whether the handbook provides appropriate considerations for identifying data sets for dose-response analysis based on systematic review conclusions. The methods described to select studies, although not always completely clear, generally align with recommendations from previous National Academies reports. For instance, the handbook indicates that, in general, toxicity values are developed when available evidence is judged either to "demonstrate" or "indicate (likely)" that a chemical causes a health effect in humans under relevant exposure circumstances. However, the handbook is not clear as to which judgments would have been made during hazard identification and which are to be made at this point in the assessment process. For instance, while strength of evidence is mentioned in the introduction of Section 12.2 of the handbook, there is no discussion of *how* the strength of evidence conclusions, or the considerations that went into drawing those conclusions, affects the selection of endpoints or studies. Additionally, some of the criteria that are described in Chapter 12 as additional factors in study selection, such as "relevance of the test species," "relevance of the studied exposure to human environmental exposures," and "quality of measurements of exposure and outcomes," would have already been considered in either the synthesis or individual study evaluations. Because the handbook does not clearly distinguish between evaluations made as part of hazard identification and considerations specific to toxicity value derivation, an analyst may be led to repeat evaluations made earlier in the assessment process.

Chapter 12 also lacks clarity on how studies should be selected for toxicity values development. The process of study selection is likely to vary on a case-by-case basis, and Chapter 12 does attempt to describe the possibilities that may arise. As judgment is used to select these studies, detailed information will need to be provided in the individual IRIS assessments in order for the process to remain transparent. The handbook states that high-quality or medium-

quality studies will be used primarily for toxicity value derivation, and yet qualifications indicate that low-quality studies may also be important, without details on when this may apply. Similar to the previous chapters, Chapter 12 is unclear as to how mechanistic data and other supporting studies should be used, or how this information should be incorporated into the assessment process. The chapter suggests that summary tables be used to document study selection and graphical comparisons be used for toxicity value selection, but no examples are provided, leaving the design up to those conducting an IRIS assessment. Specific studies are described that should not be taken through the assessment, such as non-relevant routes of exposure (excluded in PECO) and studies that do not support point of departure (POD) derivation.

Key questions about how the study selection process incorporates (1) the results of the systematic reviews that support the overall hazard conclusions and (2) additional factors not previously considered in hazard identification include the following:

- *What evidence streams are eligible to be selected?* The handbook does not address whether a minimum "summary strength of evidence judgment" (e.g., of at least "moderate") is generally required for studies within an evidence stream to be considered.
- *Which endpoints within an evidence stream are eligible?* The handbook does not address how the factors that increase/decrease certainty in the evidence (see Table 11-1 in the handbook (p. 11-5)) are incorporated into the study selection process.
- *Which individual studies that are relevant to a selected endpoint are eligible?* Although the handbook states that studies of "high or medium" confidence are preferred, it does not address how the ratings within each domain of study evaluation are incorporated into study selection.
- *What additional study characteristics are to be considered that are specific to toxicity value derivation?* The handbook describes numerous attributes in Table 12-2 (p. 12-6) that are used to evaluate studies for toxicity value derivation, but some of these (e.g., "measurement of exposure") may overlap with risk of bias or adequacy of reporting that may have been addressed earlier in the assessment. Removing from this step those aspects that have been previously evaluated as part of hazard identification would ensure that they are not "double counted" and that resources could be focused on toxicity value-specific evaluation criteria.

The handbook in many places describes what is "preferred"; however, it is not clear whether and how these attributes are tracked and integrated into confidence ratings or rankings of the resulting toxicity values. A summary table or evaluation for each study alongside these attributions would provide transparency and a basis for selection of overall toxicity values.

The committee also considered whether the basic methods for dose-response modeling and deriving toxicity values are consistent with the state of the science and presented with sufficient clarity. In Chapter 13, most of the sections appropriately refer to other EPA guidance documents, but the process of developing a toxicity value is not easy to follow. For toxicity value development, the methods are based on historical guidance documents and generally have not been updated in this handbook. The primary reference in the handbook for noncancer toxicity values, *A Review of the Reference Dose and Reference Concentration Processes* (EPA, 2002), is not actually a guidance document but a technical review of the (historical) IRIS process for developing reference concentrations (RfCs) and reference doses (RfDs). The technical panel

tasked with preparing that review was created in 1999, and the report specifically states that it "is a review, not guidance, but it does make recommendations that should be considered in the implementation of changes in the current process and/or development of needed guidance" (EPA, 2002, p. xiii). Although methods for evaluating multiple endpoints and combining multiple studies are provided, it is unclear how multiple values should be used to develop a final toxicity value, which may add to the length and complexity of some assessments. Chapter 13 would better serve users engaged in the derivation of toxicity values by citing EPA documents or other published literature that describe details of the extensive methods involved, rather than repeating the details provided in those other documents. When the handbook addresses a concept or method being updated or that is not covered elsewhere, it would need to provide enough detail for guiding a user's implementation of the approach.

Section 13.2 is confusing because it contains highly varied levels of detail and the ordering of information is hard to follow. For instance, there is a substantial amount of detail on physiologically based pharmacokinetic (PBPK) modeling and exposure route-to-route extrapolation, even though these are not routinely applied or available. Much of the detailed discussion in the chapter of procedures and approaches for those topics involving pharmacokinetic modeling could be streamlined by citing other comprehensive documents or published articles, or moving detailed discussion to an appendix.

The handbook is unclear concerning the use of probabilistic approaches for toxicity value derivation, instead of the traditional approach based on deterministic uncertainty factors. (See NRC (1994, 2009) for a discussion of the two kinds of approaches.) EPA stated that these probabilistic approaches would be routinely applied when feasible, but the handbook only makes a cursory and oblique reference to them (Thayer, 2021). The committee endorses the recommendations from previous National Academies report (NRC, 2014) for the IRIS program to transition away from using traditional deterministic approaches for deriving reference values and instead use probabilistic methods to derive risk-specific values. Since the 2014 National Academies report, several publications have further refined probabilistic approaches (e.g., Chiu and Slob, 2015; WHO & IPCS, 2018) and demonstrated their feasibility both for oral exposures (Chiu et al., 2018) and a case example for acrolein inhalation exposures (Blessinger et al., 2020).

FINDINGS AND RECOMMENDATIONS

Findings and Tier 1 Recommendations

Finding: Chapter 12 of the handbook lacks clarity about how the study selection process incorporates (1) the results of the systematic reviews that support the overall hazard conclusions and (2) additional factors not previously considered in hazard identification. Key questions include the following:

- What evidence streams are eligible to be selected?
- Which endpoints within an evidence stream are eligible?
- Which individual studies that are relevant to a selected endpoint are eligible?
- What additional study characteristics are to be considered that are specific to toxicity value derivation?

Because the handbook does not clearly distinguish between evaluations made as part of hazard identification and considerations specific to toxicity value derivation, an analyst may be led to repeat evaluations made earlier in the assessment process.

> **Recommendation 7.1:** The handbook should provide a clearer, step-by-step description of study selection, using a framework incorporating the different steps of hazard identification (including study evaluation, synthesis, and integration) as well as new steps specific to toxicity value derivation. The handbook should provide a template for how IRIS assessments are to summarize (e.g., in a table) the study selection process as applied to each endpoint, health outcome, study, and evidence stream in order to provide transparency as to study evaluation for toxicity value derivation, and to support selection of overall toxicity values. It is especially important to capture study attributes for which EPA has designated an option as "preferred" versus "less preferred." *[Tier 1]*

Finding: Chapter 13 provides important information relating to issues and considerations for deriving PODs for toxicity values, but it lacks a consistent level of detail for deriving and utilizing PODs. Most of the sections appropriately refer to other EPA guidance documents but sometimes the sections unnecessarily repeat information from those documents, rather than focus on new information to facilitate implementation of frequently used methods. There is a substantial amount of detail on PBPK modeling and route-to-route extrapolation, even though these are not routinely applied or available.

> **Recommendation 7.2:** EPA should streamline Chapter 13 of the handbook, especially Section 13.2, to focus on the most common methods and approaches rather than detailing less common scenarios. For instance, although use of PBPK modeling is designated as "preferred," it requires too much detail in this handbook to provide instructions on development and application of such models; citing other documents is preferable. If there is important information that is missing from existing EPA documents or from the peer-reviewed literature, these could be provided in an appendix to avoid disrupting the flow of the handbook. Additionally, there may be concerns over providing duplicated information, as any future updates to the related EPA documents would require an update to the handbook as well. *[Tier 1]*

Finding: The handbook is unclear as to the use of probabilistic approaches to replace the traditional deterministic uncertainty factor-based approach for toxicity value derivation.

> **Recommendation 7.3:** EPA should make it explicit in the handbook that probabilistic approaches to derive risk-specific doses will be routinely applied where feasible, referencing recent literature including a 2020 case study on acrolein (Blessinger et al., 2020. EPA should also consider when and how to transition fully away from the traditional deterministic approach to adopt risk-specific doses for its IRIS toxicity values. *[Tier 1]*

There are no Tier 2 recommendations in this chapter.

Finding and Tier 3 Recommendation

Finding: Chapter 13 of the handbook is primarily a reference chapter that directs users to various historical EPA guidance documents for deriving toxicity values, most of which have not been updated to reflect the most recently used procedures. The IRIS program and its handbook would benefit from being able to refer to updated EPA guidance.

> **Recommendation 7.4:** EPA should consider updating existing EPA guidance documents, notably those used in the development of oral RfDs and inhalation RfCs. *[Tier 3]*

References

Anglemyer, A.T., D. Krauth, and L. Bero. 2015. Industry sponsorship and publication bias among animal studies evaluating the effects of statins on atherosclerosis and bone outcomes: A meta-analysis. BMC Medical Research Methodology 15:12. 10.1186/s12874-015-0008-z.

Arzuaga, X., M.T. Smith, C.F. Gibbons, N.E. Skakkebæk, E.E. Yost, B.E.J. Beverly, A.K. Hotchkiss, R. Hauser, R.L. Pagani, S.M. Schrader, L. Zeise, and G. S. Prins. 2019. Proposed key characteristics of male reproductive toxicants as an approach for organizing and evaluating mechanistic evidence in human health hazard assessments. Environmental Health Perspectives 127:6 CID: 065001. https://doi.org/10.1289/EHP5045.

Balshem, H., M. Helfand, H.J. Schünemann, A.D. Oxman, R. Kunz, J. Brozek, G.E. Vist, Y. Falck-Ytter, J. Meerpohl, and S. Norris. 2011. GRADE guidelines: 3. Rating the quality of evidence. Journal of Clinical Epidemiology 64(4):401-406.

Becker, R.A., D.A. Dreier, M.K. Manibusan, L.A.T. Cox, T.W. Simon, and J.S. Bus. 2017. How well can carcinogenicity be predicted by high throughput "characteristics of carcinogens" mechanistic data? Regulatory Toxicology and Pharmacology 90:185-196. doi: 10.1016/j.yrtph.2017.08.021. Epub 2017 Sep 1.PMID: 28866267.

Bero, L., A. Anglemyer, H. Vesterinen, and D. Krauth. 2015. The relationship between study sponsorship, risks of bias, and research outcomes in atrazine exposure studies conducted in non-human animals: Systematic review and meta-analysis. Environmental International 92-93:597-604. http://dx.doi.org/10.1016/j.envint.2015.10.011.

Bero L, Chartres N, Diong J, Fabbri A, Ghersi D, Lam J, Lau A, McDonald S, Mintzes B, Sutton P, Turton JL. The risk of bias in observational studies of exposures (ROBINS-E) tool: concerns arising from application to observational studies of exposures. Systematic reviews. 2018 Dec;7(1):1-1.

Blessinger, T., A. Davis, W.A. Chiu, J. Stanek, G.M. Woodall, J. Gift, K.A. Thayer, and D. Bussard. 2020. Application of a unified probabilistic framework to the dose-response assessment of acrolein. Environment International 143:105953. https://doi.org/10.1016/j.envint.2020.105953.

Chartres, N., A. Fabbri, and L.A. Bero. 2016. Association of industry sponsorship with outcomes of nutrition studies: A systematic review and meta-analysis. JAMA International Medicine 176(12):1769-1777. 10.1001/jamainternmed.2016.6721.

Chiu, W.A., and W. Slob. 2015. A unified probabilistic framework for dose–response assessment of human health effects. Environmental Health Perspectives 123(12):1241-1254. https://doi.org/10.1289/ehp.1409385.

Chiu, W.A., D.A. Axelrad, C. Dalaijamts, C. Dockins, K. Shao, A.J. Shapiro, and G. Paoli. 2018. Beyond the RfD: Broad application of a probabilistic approach to improve chemical dose-response assessments for noncancer effects. Environmental Health Perspectives 126:067009. http://dx.doi.org/10.1289/EHP3368.

Downes, M.J., M.L. Brennan, H.C. Williams, and R.S. Dean. 2016. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open 6(12):e011458.

EPA (U.S. Environmental Protection Agency). 2002. A Review of the Reference Dose and Reference Concentration Processes. EPA/630/P-02/002F. U.S. Environmental Protection Agency, Risk Assessment Forum, Washington, DC. https://www.epa.gov/sites/production/files/2014-12/documents/rfd-final.pdf.

EPA. 2005. Guidelines for Carcinogen Risk Assessment. EPA/630/P-03/001B. U.S. Environmental Protection Agency, Risk Assessment Forum, Washington, DC. https://www.epa.gov/sites/production/files/2013-09/documents/cancer_guidelines_final_3-25-05.pdf.

EPA. 2018. An Umbrella Quality Assurance Project Plan (QAPP) for PBPK Models. ORD QAPP ID No: B-0030740-QP-1-1. https://hero.epa.gov/hero/index.cfm/reference/details/reference_id/4326432.

EPA. 2019a. Systematic Review Protocol for the PFBA, PFHxA, PFHxS, PFNA, and PFDA IRIS Assessments. EPA/635/R-19/049. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=345065#tab-3.

EPA. 2019b. Systematic Review Protocol for the Polychlorinated Biphenyls (PCBs) Noncancer IRIS Assessment (Preliminary Assessment Materials). EPA/635/R-19/201. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=237359.

EPA. 2020a. ORD Staff Handbook for Developing IRIS Assessments. Version 1.0, EPA/600/R-20/137. Center for Public Health and Environmental Assessment, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC.

EPA. 2020b. IRIS Assessment Plan for Oral Exposure to Vanadium and Compounds (Scoping and Problem Formulation Materials). EPA/635/R-20/112. U.S. Environmental Protection Agency, Washington, DC. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=348792.

EPA. 2020c. Systematic Review Protocol for the Methylmercury IRIS Assessment (Preliminary Assessment Materials). EPA/635/R-19/243. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=345309#tab-3.

EPA. 2021a. Vanadium Systematic Review Protocol released for public comment April 2021. Systematic Review Protocol for the Vanadium and Compounds (Oral Exposure) IRIS Assessment (Preliminary Assessment Materials). EPA/635/R-21/047. U.S. Environmental Protection Agency, Washington, DC. https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=351421.

EPA. 2021b. Systematic Review Protocol for the Inorganic Mercury Salts IRIS Assessment (Preliminary Assessment Materials). EPA/635/R-20/239. U.S. Environmental Protection Agency, Washington, DC. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=349284.

EFSA (European Food Safety Authority), U. Gundert-Remy, J. Bodin, C. Bosetti, R. FitzGerald, A. Hanberg, U. Hass, C. Hooijmans, A.A. Roney, C. Rousselle, H. van Loveren, D. Wölfle, F. Barizzone, C. Croera, C. Putzu, and A.F. Castoldi. 2017. Bisphenol A (BPA) Hazard Assessment Protocol. 10.5281/zenodo.1116881.

Friedman, L., and M. Friedman. 2016. Financial conflicts of interest and study results in environmental and occupational health research. Journal of Occupational and Environmental Medicine 58(3):238-247. 10.1097/jom.0000000000000671.

Goodman, J., and H. Lynch. 2017. Improving the International Agency for Research on Cancer's consideration of mechanistic evidence. Toxicology and Applied Pharmacology 319:39-46. 10.1016/j.taap.2017.01.020. Epub 2017 Feb 3.PMID: 28162991.

Guyatt, G.H., A.D. Oxman, R. Kunz, J. Woodcock, J. Brozek, M. Helfand, P. Alonso-Coello, P. Glasziou, R. Jaeschke, and E.A. Akl. 2011. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. Journal of Clinical Epidemiology 64(12):1294-1302.

Guyton, K.Z., I. Rusyn, W.A. Chiu, D.E. Corpet, M. van den Berg, M.K. Ross, D.C. Christiani, F.A. Beland, and M.T. Smith. 2018. Application of the key characteristics of carcinogens in cancer hazard identification. Carcinogenesis 39(4):614-622. 10.1093/carcin/bgy031. PMID: 29562322.

Haddaway, N.R., and S. Crowe. 2018. Stakeholder Engagement in Environmental Evidence Synthesis. Mistra Council for Evidence-Based Environmental Management (EviEM). https://play.google.com/store/books/details?id=13AkxQEACAAJ.

Higgins, J.P.T., and J. Thomas. 2019. Cochrane Handbook for Systematic Reviews of Interventions. Second Edition. Hoboken NJ: Wiley Blackwell.

Hill, A.B. 1965. The environment and disease: Association or causation? Proceedings of the Royal Society of Medicine 58:295-300.

Holman, B., L. Bero, B. Mintzes. 2019. Industry Scholarship bias. Catalogue of Bias. https://catalogofbias.org/biases/industry-sponsorship-bias/

Hooijmans, C.R., R.B.M. de Vries, M. Ritskes-Hoitinga, M.M. Rovers, M.M. Leeflang, J. IntHout, K.E. Wever, L. Hooft, H. de Beer, T. Kuijpers, M.R. Macleod, E.S. Sena, G. Ter Riet, R.L. Morgan, K.A. Thayer, A.A. Rooney, G.H. Guyatt, H.J. Schünemann, and M.W. Langendam. 2018. Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. PLoS One 13(1):e0187271.

Huss, A., M. Egger, K. Hug, K. Huwiler-Müntener, and M. Röösli. 2007. Source of funding and results of studies of health effects of mobile phone use: Systematic review of experimental studies. Environmental Health Perspectives 115(1):1-4. 10.1289/ehp.9149.

IARC (International Agency for Research on Cancer). 2019. Preamble to the IARC Monographs (amended January 2019). https://monographs.iarc.who.int/wp-content/uploads/2019/07/Preamble-2019.pdf.

Igelström E, Campbell M, Craig P, Katikireddi SV. Cochrane's risk-of-bias tool for non-randomised studies (ROBINS-I) is frequently misapplied: a methodological systematic review: ROBINS-I frequently misapplied: systematic review. Journal of Clinical Epidemiology. 2021 Aug 23.

IOM (Institute of Medicine). 2011. Finding What Works in Health Care: Standards for Systematic Reviews. Washington, DC: The National Academies Press. https://doi.org/10.17226/13059.

James, K.L., N.P. Randall, and N.R. Haddaway. 2016. A methodology for systematic mapping in environmental sciences. Environmental Evidence 5(1). 10.1186/s13750-016-0059-6.

Keshava, C., J.A. Davis, J. Stanek, K.A. Thayer, A. Galizia, N. Keshava, J. Gift, S.V. Vulimiri, G. Woodall, C. Gigot, K. Garcia, A. Greenhalgh, B. Schulz, S. Volkoff, K. Camargo, and A.S. Persad. 2020. Application of systematic evidence mapping to assess the impact of new research when updating health reference values: A case example using acrolein. Environment International 143:105956. 10.1016/j.envint.2020.105956.

La Merrill, M.A., L.N. Vandenberg, M.T. Smith, W. Goodson, P. Browne, H.B. Patisaul, K.Z. Guyton, A. Kortenkamp, V.J. Cogliano, T.J. Woodruff, L. Rieswijk, H. Sone, K.S. Korach, A.C. Gore, L. Zeise, and R.T. Zoeller. 2020. Consensus on the key characteristics of endocrine-disrupting chemicals as a basis for hazard identification. Nature Reviews Endocrinology 16:45-57. https://doi.org/10.1038/s41574-019-0273-8.

Lind, L., J.A. Araujo, A. Barchowsky, S. Belcher, B.R. Berridge, N. Chiamvimonvat, W.A. Chiu, V.J. Cogliano, S. Elmore, A.K. Farraj, A.V. Gomes, C.M. McHale, K.B. Meyer-Tamaki, N.G. Posnack, H.M. Vargas, X. Yang, L. Zeise, C. Zhou, and M.T. Smith. 2021. Key characteristics of cardiovascular toxicants. Environmental Health Perspectives 129(9):95001. 10.1289/ehp9321.

Luderer, U., B. Eskenazi, R. Hauser, K.S. Korach, C.M. McHale, F. Moran, L. Rieswijk, G. Solomon, O. Udagawa, L. Zhang, M. Zlatnik, L. Zeise, and M.T. Smith. 2019. Proposed key characteristics of female reproductive toxicants as an approach for organizing and evaluating mechanistic data in hazard assessment. Environmental Health Perspectives 127(7): 075001-1–075001-14. https://doi.org/10.1289/EHP4971.

Lundh, A., J. Lexchin, B. Mintzes, J.B. Schroll, and L. Bero. 2017. Industry sponsorship and research outcome. Cochrane Database of Systematic Reviews 2:Mr000033. 10.1002/14651858.MR000033.pub3.

Lundh, A., K. Rasmussen, L. Østengaard, I. Boutron, L.A. Stewart, and A. Hróbjartsson. 2019. Systematic review finds that appraisal tools for medical research studies address conflicts of interest superficially. Journal of Clinical Epidemiology 120:104-115.

Michaels D. 2006. Manufactured uncertainty: protecting public health in the age of contested science and product defense. Ann NY Acad Sci. 1076:149-62. 10.1196/annals.1371.058

Minozzi S, M. Cinquini, S. Gianola, G. Castellini, C. Gerardi, R. Banzi. 2019. Risk of bias in nonrandomized studies of interventions showed low inter-rater reliability and challenges in its application. Journal of Clinical Epidemiology 112:28-35.

Moga, C., B. Guo, D. Schopflocher, and C. Harstall. 2012. Development of a Quality Appraisal Tool for Case Series Studies Using a Modified Delphi Technique. Edmonton, AB: Institute of Health Economics. https://www.ihe.ca/advanced-search/development-of-a-quality-appraisal-tool-for-case-series-studies-using-a-modified-delphi-technique.

Morgan, R., J. Sterne, J. Higgins, K. Thayer, H. Schunemann, A. Rooney, and K. Taylor. 2017. A new instrument to assess Risk of Bias in Non-randomised Studies of Exposures (ROBINS-E): Application to studies of environmental exposure. Abstracts of the Global Evidence Summit, Cape Town, South Africa. Cochrane Database of Systematic Reviews (9 Suppl 1). https://doi.org/10.1002/14651858.CD201702.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2017a. Application of Systematic Review Methods in an Overall Strategy for Evaluating Low-Dose Toxicity from Endocrine Active Chemicals. Washington, DC: The National Academies Press. https://doi.org/10.17226/24758.

NASEM. 2017b. Using 21st Century Science to Improve Risk-Related Evaluations. Washington, DC: The National Academies Press. https://doi.org/10.17226/24635.

NASEM. 2018. Progress Toward Transforming the Integrated Risk Information System (IRIS) Program: A 2018 Evaluation. Washington, DC: The National Academies Press. https://doi.org/10.17226/25086.

NASEM. 2021. The Use of Systematic Review in EPA's Toxic Substances Control Act Risk Evaluations. Washington, DC: The National Academies Press. https://doi.org/10.17226/25952.

Nosek, B.A., and D. Lakens. 2014. Registered reports: A method to increase the credibility of published results. Social Psychology 45(3):137–141. 10.1027/1864-9335/a000192.

NRC (National Research Council). 1994. Science and Judgment in Risk Assessment. Washington, DC: The National Academies Press.https://doi.org/10.17226/2125.

NRC. 2009. Science and Decisions: Advancing Risk Assessment. Washington, DC: The National Academies Press.https://doi.org/10.17226/12209.

NRC. 2011. Review of the Environmental Protection Agency's Draft IRIS Assessment of Formaldehyde. Washington, DC: The National Academies Press. https://doi.org/10.17226/13142.

NRC. 2014. Review of EPA's Integrated Risk Information System (IRIS) Process. Washington, DC: The National Academies Press. https://doi.org/10.17226/18764.

NTP (National Toxicology Program) 2018a. Report on Carcinogens Monograph on Antimony Trioxide. Research Triangle Park, NC: U.S. Department of Health and Human Services, Public Health Service.

NTP. 2018b. Report on Carcinogens Monograph on Haloacetic Acids Found as Water Disinfection By-Products. Research Triangle Park, NC: U.S. Department of Health and Human Services, Public Health Service.

NTP. 2019. Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration. U.S Department of Health and Human Services; Office of Health Assessment and Translation. https://ntp.niehs.nih.gov/whatwestudy/assessments/noncancer/handbook/index.html.

Radke, E.G., E.E. Yost, N. Roth, S. Sathyanarayana, and P. Whaley. 2020. Application of US EPA IRIS systematic review methods to the health effects of phthalates: Lessons learned and path forward. Environment International 145:105820. 10.1016/j.envint.2020.105820.

Rezende, L.F.M., J.P. Rey-López, T.H. Sá, N. Chartres, A. Fabbri, L. Powell, E. Stamatakis, and L. Bero. 2018. Reporting bias in the literature on the associations of health-related behaviors and statins with cardiovascular disease and all-cause mortality. PLoS Biology 16(6):e2005761. https://doi.org/10.1371/journal.pbio.2005761.

Rusyn, I., X. Arzuaga, R.C. Cattley, J.C. Corton, S.S. Ferguson, P. Godoy, K.Z. Guyton, N. Kaplowitz, S.R. Khetani, R.A. Roberts, R.A. Roth, and M.T. Smith. 2021. Key characteristics of human hepatotoxicants as a basis for identification and characterization of the causes of liver toxicity. Hepatology. https://doi.org/10.1002/hep.31999.

Samet, J.M., W.A. Chiu, V. Cogliano, J. Jinot, D. Kriebel, R.M. Lunn, F.A. Beland, L. Bero, P. Browne, L. Fritschi, J. Kanno, D.W. Lachenmeier, Q. Lan, G. Lasfargues, F. Le Curieux, S. Peters, P. Shubat, H. Sone, M.C. White, J. Williamson, M. Yakubovskaya, J. Siemiatycki, P.A. White, K.Z. Guyton, M.K. Schubauer-Berigan, A.L. Hall, Y. Grosse, V. Bouvard, L. Benbrahim-Tallaa, F. El Ghissassi, B. Lauby-Secretan, B. Armstrong, R. Saracci, J. Zavadil, K. Straif, and C.P. Wild. 2020. The IARC Monographs: Updated procedures for modern and transparent evidence synthesis in cancer hazard identification. Journal of the National Cancer Institute 112(1):30-37. 10.1093/jnci/djz169.

Schaefer H.R., J.L. Myers. 2017. Guidelines for performing systematic reviews in the development of toxicity factors. Regulatory Toxicology and Pharmacology 91:124-41.

SIGN (Scottish Intercollegiate Guidelines Network). 2011. A Guideline Developer's Handbook. https://www.sign.ac.uk/assets/sign50_2011.pdf.

Smith, M.T., K.Z. Guyton, C.F. Gibbons, J.M. Fritz, C.J. Portier, I. Rusyn, D.M. DeMarini, J.C. Caldwell, R.J. Kavlock, P.F. Lambert, S.S. Hecht, J.R. Bucher, B.W. Stewart, R.A. Baan, V.J. Cogliano, and K. Straif. 2016. Key characteristics of carcinogens as a basis for organizing data on mechanisms of carcinogenesis. Environmental Health Perspectives 124(6):713-721. 10.1289/ehp.1509912. Epub 2015 Nov 24. PMID: 26600562; PMCID: PMC4892922.

Smith C.J., T.A. Perfetti, A.W. Hayes, S.C. Berry, J.E. Trosko, J.A. King, J.I. Goodman, C.G. Begley, A. Dayan. 2021. Categorizing the characteristics of human carcinogens: a need for specificity. Arch Toxicol. 95(8):2883-2889. doi: 10.1007/s00204-021-03109-w. Epub 2021 Jun 20. PMID: 34148101.

Sterne, J.A.C., M.A. Hernán, B.C. Reeves, J. Savović, N.D. Berkman, M. Viswanathan, D. Henry, D.G. Altman, M.T. Ansari, I. Boutron, J.R. Carpenter, A.W. Chan, R. Churchill, J.J. Deeks, A. Hróbjartsson, J. Kirkham, P. Jüni, Y.K. Loke, T.D. Pigott, C.R. Ramsay, D. Regidor, H.R. Rothstein, L. Sandhu, P.L. Santaguida, H.J. Schünemann, B. Shea, I. Shrier, P. Tugwell, L. Turner, J.C. Valentine, H. Waddington, E. Waters, G.A. Wells, P.F. Whiting, and J.P.T. Higgins. 2016. ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. British Medical Journal 355:i4919.

Thayer, K. 2021. Perspectives on EPA's IRIS Assessment Handbook. Presented to the Committee to Review EPA's IRIS Assessment Handbook, April 16.

van der Naald, M., S. Wenker, P.A. Doevendans, K.E. Wever, and S.A.J. Chamuleau. 2020. Publication rate in preclinical research: A plea for preregistration. BMJ Open Science 4:e100051. 10.1136/ bmjos-2019-100051.

Walker, V.R., A.L. Boyles, K.E. Pelch, S.D. Holmgren, A.J. Shapiro, C.R. Blystone, M.J. Devito, R.R. Newbold, R. Blain, P. Hartman, K.A. Thayer, and A.A. Rooney. 2018. Human and animal evidence of potential transgenerational inheritance of health effects: An evidence map and state-of-the-science evaluation. Environment International 115: 48-69. 10.1016/j.envint.2017.12.032.

Wells, E. 2017. Evidence regarding the impact of conflicts of interest on environmental and occupational health research. Current Environmental Health Reports 4:109-118. 10.1007/s40572-017-0139-y.

Whaley P., E. Aiassa, C. Beausoleil, A. Beronius, G. Bilotta, A. Boobis, R. de Vries, A. Hanberg, S. Hoffmann, N. Hunt, C.F. Kwiatkowsk, J. Lam, S. Lipworth, O. Martin, N. Randall, L. Rhomberg, A. A. Rooney, H. J. Schünemann, D. Wikoff, T. Wolffe, C. Halsall. Recommendations for the conduct of systematic reviews in toxicology and environmental health research (COSTER). Environment International 143:105926. https://doi.org/10.1016/j.envint.2020.105926.

WHO & IPCS (World Health Organization & International Programme on Chemical Safety). 2018. Guidance Document on Evaluating and Expressing Uncertainty in Hazard Characterization, 2nd ed. World Health Organization. https://apps.who.int/iris/handle/10665/259858. License: CC BY-NC-SA 3.0 IGO.

Wolffe, T.A.M., P. Whaley, C. Halsall, A.A. Rooney, and V.R. Walker. 2019. Systematic evidence maps as a novel tool to support evidence-based decision-making in chemicals policy and risk management. Environment International 130:104871. https://doi.org/10.1016/j.envint.2019.05.065.

Woodruff, T.J., P. Sutton, and The Navigation Guide Work Group. 2011. An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences. Health Affairs 30(5):931-937.

Woodruff, T.J., and P. Sutton. 2014. The Navigation Guide systematic review methodology: A rigorous and transparent method for translating environmental health science into better health outcomes. Environmental Health Perspectives 122(10):1007-1014.

Appendix A

Committee Member Biosketches

Lisa A. Bero *(Chair),* is a professor in the School of Public Health and the School of Medicine (General Internal Medicine) at the University of Colorado Anschutz Medical Center. She is also the chief scientist at the Center for Bioethics and Humanities at that medical center. In addition, she is an affiliated professor at the Charles Perkins Centre and School of Pharmacy, Faculty of Medicine and Health at the University of Sydney. Dr. Bero is an adjunct professor in the Department of Clinical Pharmacy and Institute for Health Policy Studies at the University of California, San Francisco. She is recognized for her methodological studies on bias (including publication/reporting, design, and funding biases) in the fields of clinical medicine (pharmaceuticals), tobacco control, and environmental research, and on the use and implications of the evidence for prescribing decisions/policy. She investigates hidden biases in the design, conduct, and publication of research. For more than 20 years, Dr. Bero has been actively involved in the Cochrane Collaboration, a global organization that summarizes the best evidence from research to help make informed choices about health care. She served as a member of the National Academies Board on Health Care Services; Committee to Review the IRIS Process; and Committee on Conflicts of Interest in Medical Research, Education, and Practice. Dr. Bero received a PhD in pharmacology from Duke University.

Hugh A. Barton is an independent consultant for applications of systems pharmacology and toxicology to drug discovery and safety evaluation or environmental risk assessment. He provides expert advice on physiologically based pharmacokinetic (PBPK) and pharmacodynamic (PD) models to address low-dose, interspecies, and inter-route extrapolations in estimating risks and their implementation for decision making. Dr. Barton formerly was an associate research fellow with Biomedicine Design at Pfizer, Inc. for 10 years. He focused on drug discovery by applying translational modeling and simulation to oncology, cardiovascular disease, and neurodegenerative diseases to assess pharmacokinetic PD, and safety. Prior to that, he worked for the U.S. Environmental Protection Agency (EPA) for 9 years and several consulting companies. He was a member of the National Academies Committee on Inorganic Arsenic and the Committee to Evaluate the IRIS Protocol for Inorganic Arsenic. Dr. Barton currently serves as a member of EPA's Science Advisory Board (SAB), and he previously served as chair of the SAB Chemical Assessment Advisory Committee. He received a PhD in toxicology from the Massachusetts Institute of Technology.

Weihsueh A. Chiu is a professor in the Department of Veterinary Integrative Biosciences at the Texas A&M University. His research focuses on the development of quantitative, data-driven approaches for understanding and predicting the human health effects of environmental chemicals. Specifically, his research applies computational and statistical methods to transform data into knowledge used to protect public health. He also has an interest in approaches to estimate the variability in individual susceptibility to environmental exposures, so as to better

protect sensitive subpopulations. Dr. Chiu currently serves on the National Academies Committee on Use of Emerging Science for Environmental Health Decisions. His previous service on National Academies committees includes the Committee on Endocrine-Related Low Dose Toxicity and the Committee on Predictive-Toxicology Approaches for Military Assessments of Acute Exposures. He received a PhD in physics from Princeton University.

Gary L. Ginsberg is director of the Center for Environmental Health within the New York State Department of Health and has a clinical professor appointment at the Yale School of Public Health. Previously, he was a state toxicologist in the Connecticut Department of Public Health Division of Environmental and Occupational Health Assessment. Dr. Ginsberg is involved in the use of toxicology and risk-assessment principles to evaluate human exposure to chemicals in air, water, soil, food, and the workplace. His published work includes the development and evaluation of physiologically based pharmacokinetic modeling for assessing risks from exposure to environmental agents, including neurotoxic effects, the interaction of lead and psycho-social stress, and developmental aspects of children. He served on several National Academies committees, including the Committee on Use of Emerging Science for Environmental Health Decisions, Committee on Inorganic Arsenic, Committee on Improving Risk Analysis Approaches Used by the U.S. EPA, and Committee on Human Biomonitoring for Environmental Toxicants. Dr. Ginsberg received a PhD in toxicology from the University of Connecticut.

Julie B. Herbstman is an associate professor in the Department of Environmental Health Sciences, director of the Columbia Center for Children's Environmental Health, and co-director of the Certificate Program in Molecular Epidemiology at the Columbia University Mailman School of Public Health. Her recent research involves the integration of epigenetic biomarkers to explore the mechanistic pathway between prenatal exposures and disease risk. In addition, she has addressed the impact of prenatal exposures to environmental pollutants, including polybrominated diphenyl ethers and polycyclic aromatic hydrocarbons on child growth and development. She has also been involved in research exploring the long-term environmental health impact of exposure to pollutants from the collapse of the World Trade Center on 9/11. Dr. Herbstman received a PhD in environmental epidemiology from the Johns Hopkins University.

Jessica L. Myers is a toxicologist and risk assessor. She is currently working at the Texas Commission on Environmental Quality (TCEQ). Dr. Myers co-authored TCEQ guidance on conducting systematic reviews during the development of chemical-specific toxicity factors. She currently serves on the National Academies Committee to Review EPA's TSCA Systematic Review Guidance Document. She has a PhD in cell and molecular biology from The University of Texas at Austin.

Heather B. Patisaul is the associate dean for research in the College of Sciences and a professor in the Department of Biological Sciences at North Carolina State University. She explores the mechanisms by which endocrine disrupting compounds alter neuroendocrine pathways in the brain related to sex specific physiology and behavior. She is specifically interested in phytoestrogens, flame retardants, and BPA (Bisphenol A). Dr. Patisaul is a National Institute of Environmental Health Sciences Outstanding New Environmental Scientist award recipient (2007) and has participated on several national and international expert panels and workshops related to health effects associated with soy and other endocrine disruptors. She chaired the 2016

Gordon Research Conference on Environmental Endocrine Disruptors, and has co-edited several special issues on endocrine disruptors, brain, and behavior. In addition, Dr. Patisaul served on three previous National Academies committees: the workshop planning Committee on Understanding the Paradigm Change at the Interface of Emerging Sources of Environmental Health Data and Decision Making, Committee on Incorporating 21st Century Science in Risk-Based Evaluations, and Committee to Review EPA's Draft State of the Science Paper on Nonmonotonic Dose Response. She received a PhD in population biology, ecology, and evolution from Emory University.

David B. Richardson is a professor in the Department of Epidemiology, School of Public Health at the University of North Carolina at Chapel Hill. He is also deputy director of the North Carolina Occupational Safety and Health Education and Research Center and director of the center's Program in Occupational Epidemiology. His research focuses on the health effects of occupational and environmental exposures, particularly with regard to carcinogens. He has conducted studies of cancer among workers in the United States and abroad. Dr. Richardson's current research includes studies of mortality among nuclear industry workers and uranium miners, and development of innovative methods for occupational cancer studies. He is an associate editor of the journals *Occupational and Environmental Medicine*, *American Journal of Epidemiology,* and *Environmental Health Perspectives*. His service on National Academies committees includes the Committee to Review the Health Effects in Vietnam Veterans of Exposure to Herbicides—Tenth Biennial Update and the Committee on the Review of the Department of Labor's Site Exposure Matrix Database. Dr. Richardson received a PhD and MSPH, both in epidemiology, from the University of North Carolina at Chapel Hill.

Paul Whaley is a researcher and academic editor specializing in systematic review methods for environmental health research and chemical risk assessment. He is based at Lancaster Environment Centre, at Lancaster University, UK. Mr. Whaley is also a research fellow of the Evidence-based Toxicology Collaboration (EBTC) at the Johns Hopkins Bloomberg School of Public Health, where he co-leads several EBTC initiatives on improving the quality of published toxicological and environmental health research. In general, his work focuses on developing frameworks for systematic mapping and review of scientific evidence in support of assessing and managing human health risks posed by exposure to environmental challenges. Mr. Whaley is systematic reviews editor for *Environment International*. He has an MLitt in philosophy from the University of St Andrews and is pursuing a PhD in research synthesis methods in environmental health at Lancaster University.

Kristine L. Willett is a professor of pharmacology and environmental toxicology and chair of the Department of BioMolecular Sciences at the University of Mississippi. Her research interests include using fish models to study mechanisms of polycyclic aromatic hydrocarbon and cannabinoid toxicity as well as in new drug discovery for cancer and epilepsy. Dr. Willett also has studied nanosilver mechanisms of toxicity and the consequences of the Deep Water Horizon Oil Spill on oysters. She has led research projects that were designed to fundamentally understand the molecular mechanisms underlying toxicity and/or shed light on the potential adverse outcomes due to relevant anthropogenic contamination. She is a deputy editor for *Toxicological Sciences*. Dr. Willett received a PhD in toxicology from Texas A&M University.

Corwin M. Zigler joined the faculty of The University of Texas at Austin in 2018, sharing joint appointments in the Department of Statistics and Data Sciences and the Department of Women's Health at Dell Medical School. Prior to joining the University of Texas, he was faculty in the Department of Biostatistics at the Harvard T.H. Chan School of Public Health. Dr. Zigler's primary statistical research focus is Bayesian statistical methods for making causal inferences with observational data. His work has focused on problems in environmental health and environmental policy, with key contributions in air pollution epidemiology, regulatory policy, studies of point-source exposure, and health impacts of power plant emissions. His work integrates methods from across the disciplines of statistics, epidemiology, and atmospheric science and engineering. Specific areas of application to environmental health include evaluation of federal nonattainment designations for national ambient air quality standards and evaluation of strategies to control harmful pollution emissions from power plants. He currently serves as associate editor for the journals *Biometrics* and *Biostatistics*, and is heavily involved through elected positions in the Health Policy Statistics Section of the American Statistical Association. His career awards include the 2010 Carolbeth Korn Prize for the most outstanding graduating student in the University of California, Los Angeles School of Public Health, a 2012 Young Investigator Award from the Statistics in Epidemiology Section of the American Statistical Association, and the 2019 Rothman Prize for the best paper published in *Epidemiology*. Dr. Zigler received a PhD in biostatistics from the University of California, Los Angeles.

Appendix B

2020 IRIS Handbook Organization and Contents and EPA Questions to the Committee

2020 HANDBOOK ORGANIZATION AND CONTENTS[a]

Assessment Development Stage	Chapter	Summary Description
Scoping	1	Defines the parameters of the assessment based on EPA needs.
Problem formulation and IRIS assessment plan (IAP) development	2	Describes health effects of potential interest and key science issues.
Systematic review protocol	3	Systematic review procedures for populations, exposures, comparators, and outcomes (PECO) criteria; literature identification; study evaluation; and data extraction.
Literature search, screening, and inventory	4	Describes methods for performing comprehensive literature search(es). Uses PECO criteria to identify relevant human and animal health effect studies. Identifies absorption, distribution, metabolism, and excretion (ADME) studies; models; and mechanistic information. Categorizes studies (e.g., by study type, health effect) and extracts cursory information to allow for organization by study design/mechanism.
Refined evaluation plan	5	Describes process for deciding whether and how to prioritize and group sets of related endpoints into health effect categories for review, focusing on those most likely to inform hazard identification.
Study evaluation	6	Describes study evaluation methods for individual human and animal health effect studies, pharmacokinetic models, and an approach for mechanistic studies. Study evaluation includes consideration of reporting quality, risk of bias, and sensitivity.
Organize hazard review	7	Discusses approaches to finalize the utility and organization of health effect categories and studies for hazard identification. These decisions are informed by study evaluation, toxicokinetic, and consideration of mechanistic information.
Data extraction and display	8	Presents types of key health effect study information to collect in a database and examples of graphical and tabular displays.
Evidence synthesis • Human and animal studies	9	Discusses considerations and approaches to analyze results incorporating the strengths and limitations of the sets of health effect studies of exposed humans (controlled exposure or epidemiology) and animal toxicology experiments by health effect or other grouping.

• Mechanistic information	10	Presents a process to conduct focused, step-wise analyses of the most relevant mechanistic evidence and summarize results by health effect or other grouping based on the unique needs of the assessment (e.g., key science issues) and considerations that arise from analyzing the human and animal evidence (e.g., questions of biological plausibility or human relevance).
Evidence integration	11	Describes the contents of the evidence integration narrative for hazard identification and a framework to determine overall summary conclusions.
Hazard considerations and study selection for deriving toxicity values	12	Describes the selection process to determine the most informative studies and outcomes for dose-response analysis based on study confidence and other considerations including hazard judgments and susceptibility.
Derive toxicity values	13	Describes dose-response modeling and methods to develop a quantitative estimate for each hazard of concern (cancer and noncancer). This includes the consideration of uncertainty and susceptibility and description of confidence in the estimates.

[a] https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=350086. Under "Downloads," see "Draft Charge Questions for Reviewers."

QUESTIONS FROM EPA TO THE COMMITTEE

Peer review advice on the following charge questions will be most useful when prioritized to indicate its relative importance during revision:

- *Tier 1: Recommended Revisions*—Highest priority recommendations the committee believes are critical to improve the scientific rigor and/or clarity of the document.
- *Tier 2: Suggestions*—Recommendations that EPA should consider to strengthen the document.
- *Tier 3: Future Considerations*—Topic areas that may inform future developments. These recommendations are outside the immediate scope and/or needs of the current document under review.

Please comment on each question below, elaborating on the rationale and scientific evidence relating to each comment, and do not limit comments to "yes" or "no." For Tier 1 and Tier 2 recommendations, please provide specific revisions or alternatives to improve the clarity of the presentation and increase the scientific rigor of the approach.

1. Please comment on the overall organization of the handbook, in particular on whether the key aspects of the assessment process are represented.

2. Are the systematic review approaches used by the IRIS Program (outlined in Chapters 1-5) clearly described and consistent with methodologies considered to be state of the science by experts in the field?

3. Are the study evaluation methods in Chapter 6 for individual human studies (epidemiology and controlled exposure), animal studies, mechanistic evidence (pilot testing approaches), and pharmacokinetic models adequate; if not, how can they be improved (Chapter 6)?

4. Given the broad questions considered in IRIS assessments (typically necessitating multiple systematic reviews and dose-response analyses to address different health effects, exposure scenarios, and potential susceptible populations or life stages), the handbook outlines approaches for refinement of the scope and analyses in the assessment. This relates to multiple stages of assessment development, primarily problem formulation (Chapters 2), inventorying the literature (Chapter 4), refinement of the evaluation plan (Chapter 5), and organizing the hazard review (Chapter 7). These processes are used to inform a variety of subsequent assessment decisions, such as which health outcomes to focus on during the hazard review, how to approach the evaluation of mechanistic evidence, and identifying scientific complexities related to application of pharmacokinetic models or dose-response analysis. Does the handbook clearly lay out a state-of-the-science approach to refinement? Are there specific areas for improvement (please indicate recommended alternatives)?

5. Sections 2.2, 4.3.3, and 6.6 and Chapter 10 describe the systematic process for evaluating mechanistic data. Please provide your review and assessment of the handbook's process for evaluating and integrating mechanistic data. Are there specific areas for improvement (please indicate recommended alternatives)?

6. Chapters 9-11 of the handbook outline approaches for applying expert judgment to synthesize and integrate the available evidence based on considerations related to conclusions about the likelihood of a biologically plausible causal relationship. Are the approaches to evidence synthesis described in Chapters 9 and 10 scientifically sound? Are the considerations sufficiently broad to allow for application to the wide range of scenarios that will be encountered when applied to individual assessments? Are the methods sufficiently clear in describing the intent to synthesize the relevant evidence, incorporating study evaluation conclusions, regardless of the study results? Are the approaches described in Chapter 11 scientifically sound and appropriate for integrating the various types of evidence relevant to investigating the potential for human health effects from exposure to environmental chemicals?

7. Chapter 11 presents five categories for drawing evidence integration conclusions, which builds upon judgments regarding the available human, animal, and mechanistic evidence. This approach was previously reviewed by the NAS in NASEM (2018) and presented in systematic review protocols released for public comment during 2018 and 2019[1]. More recently, the IRIS Program considered a three-category approach for evidence integration that was disseminated in systematic review protocols released for

[1] The IRIS systematic review protocols released for public comment with the five-category approach included EPA (2018, 2019a,b).

public comment in 2019 and 2020.[2] Please comment specifically on which of these approaches is recommended, why it is recommended, and any specific refinements for improvement.

8. Chapters 12 and 13 outline considerations and approaches for selecting studies, specific health effects, and endpoints for dose-response analysis, and for selecting toxicity values. They also provide an overview of methods for conducting dose-response modeling and deriving toxicity values based on more detailed, existing guidance and recommendations. Does the handbook provide appropriate considerations for identifying data sets for dose-response analysis based on systematic review conclusions? Are the basic methods for dose-response modeling and deriving toxicity values consistent with the current state of the science, and presented with sufficient clarity?

REFERENCES

EPA (U.S. Environmental Protection Agency). 2018. Systematic Review Protocol for the IRIS Chloroform Assessment (Inhalation) (Preliminary Assessment Materials). U.S. Environmental Protection Agency, Washington, DC, EPA/635/R-17/486. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=338653.

EPA. 2019a. Systematic Review Protocol for the Hexavalent Chromium IRIS Assessment (Preliminary Assessment Materials). U.S. Environmental Protection Agency, Washington, DC, EPA/635/R-18/155. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=343950#tab-3.

EPA. 2019b. Updated Problem Formulation and Systematic Review Protocol for the Inorganic Arsenic IRIS Assessment. U.S. Environmental Protection Agency, Washington, DC, EPA/635/R-19/049. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=343951.

EPA. 2019c. Systematic Review Protocol for the PFAS IRIS Assessments. U.S. Environmental Protection Agency, Washington, DC, EPA/635/R-19/049. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=345065.

EPA. 2019d. Systematic Review Protocol for the Polychlorinated Biphenyls (PCBs) Noncancer IRIS Assessment (Preliminary Assessment Materials). U.S. Environmental Protection Agency, Washington, DC, EPA/635/R-19/201. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=237359.

EPA. 2020. Systematic Review Protocol for the Methylmercury (MeHg) IRIS Assessment (Preliminary Assessment Materials). U.S. Environmental Protection Agency, Washington, DC, EPA/635/R-19/243. https://cfpub.epa.gov/ncea/iris_drafts/recordisplay.cfm?deid=345309.

NASEM (National Academies of Sciences, Engineering, and Medicine). 2018. Progress Toward Transforming the Integrated Risk Information System (IRIS) Program: A 2018 Evaluation. Washington, DC: The National Academies Press. https://doi.org/10.17226/25086.

[2] The IRIS systematic review protocols released for public comment with the three-category approach included EPA (2019c,d, 2020).

Appendix C

Open Session Agendas

February 11, 2021
Web Meeting

11:00 AM (EST)	Opening Remarks Lisa Bero, Committee Chair
11:05	IRIS Assessments Prior to the Development of the Current IRIS Handbook Jonathan Samet, Colorado School of Public Health
	Questions from Committee Members
12:00 PM	Opportunity for Public Comment
12:20	End of Public Session

April 16, 2021
Web Meeting

10:00 AM (EDT)	Welcome Lisa Bero, Committee Chair
10:05	Perspectives on EPA's IRIS Assessment Handbook Kris Thayer (EPA)
10:50	Committee Discussion with EPA Personnel Kris Thayer, Xabier Arzuaga, Allen Davis, Laura Dishaw, Catherine Gibbons, Barbara Glenn, Karen Hogan, Dustin Kapraun, Andrew Kraft, Elizabeth Radke, Andy Shapiro, George Woodall, and Erin Yost
11:50	Opportunity for Public Comment
12:50 PM	End of Open Session